View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by ScholarShip

Abstract

Genome-wide identification and expression analysis of TCP transcription factors in cotton

by Jun Ma

May, 2015

Director: Dr. Baohong Zhang

DEPARTMENT OF BIOLOGY

TCP proteins are plant-specific transcription factors known to perform a variety of physiological functions during plant growth and development. In the current study, we performed for the first time the comprehensive analysis of TCP gene family in two sequenced cotton species, *Gossypium raimondii* and *Gossypium arboreum*, including phylogenetic analysis, chromosome location, gene duplication status, gene structure and conserved motif analysis, as well as expression profiles in different tissues and at different developmental stages. Our results showed that a total of 38 non-redundant cotton TCP encoding genes were identified in *G. raimondii*, unevenly distributing across 11 out of the 13 chromosomes, whereas *G. arboreum* contains 36 TCP genes, which

distributed across all of the thirteen chromosomes. GrTCPs and GaTCPs within the same subclade of the phylogenetic tree shared similar exon/intron organization and motif composition. In addition, both segmental duplication and whole-genome duplication contributed significantly to the expansion of cotton TCPs. Moreover, most TCP genes exhibited tissue-specific expression profiles, which shed light on their functional divergence. Remarkably, many these TCP transcription factor genes are specifically expressed in cotton fiber during different developmental stages, including cotton fiber initiation and early development. This suggests that TCP genes may play important roles in cotton fiber development.

Genome-wide identification and expression analysis of TCP transcription factors in cotton

A Thesis

Presented To

The Faculty of the Department of Biology

East Carolina University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Molecular Biology and Biotechnology

by

Jun Ma

May, 2015

©Copyright 2015
Jun Ma

Genome-wide identification and expression analysis of TCP transcription factors in cotton

by

Jun Ma

APPROVED BY:

DIRECTOR OF THESIS:_____
Baohong Zhang, PhD

COMMITTEE MEMBER:_____
Xiaoping Pan, PhD

COMMITTEE MEMBER:_____
Yiping Qi, PhD

COMMITTEE MEMBER:_____
Michael Brewer, PhD

COMMITTEE MEMBER:_____
Shouquan Huo, PhD

CHAIR OF THE DEPARTMENT OF BIOLOGY:

_____
Jeff McKinnon, PhD

DEAN OF THE GRADUATE SCHOOL:

_____
Paul J. Gemperline, PhD

ACKNOWLEDGEMENTS

I would like to acknowledge the advice and guidance of my advisor, Dr. Baohong Zhang, whose advice and guidance have been essential from the beginning through the culmination of this project. The time spent in his laboratory has been vital to my scientific growth on so many levels. I would also like to thank my thesis committee members, Dr. Xiaoping Pan, Dr. Yiping Qi, Dr. Michael Brewer and Dr. Shouquan Huo for their guidance and feedback.

In addition, I would like to thank the members of the Zhang's lab, especially Fuliang Xie and Faten Taki, for their help and guidance in my project. I would like to thank my family members for supporting and encouraging me to pursue this degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

Nature of the Problem and Background

Regulation of gene transcription, one of the most complex activities in cells, plays a significant role in a wide variety of biological processes, such as cell growth, cell cycle control, signal transduction, metabolic and physiological balance, and response to environmental stimuli [1-5]. Among many mechanisms of transcriptional regulation of gene expression, transcription factors are considered to be the most important. Transcription factors are a diverse family of regulatory proteins with specific DNA-binding domains involved in the regulation of many cellular processes by either stimulating or repressing transcription of the related genes [2,3,6]. Up to now, more than 60 transcription factor families have been identified in plants [7].

The TCP proteins are a family of transcription factors exclusive to higher plants and involved in the regulation of cell growth and proliferation [8,9]. This class of transcription factors are featured by a highly conserved 59-residue-long DNA-binding motif at the N-terminus called TCP domain, which is named after its founding members: TB1 (TEOSINTE BRANCHED 1) in *Zea mays*, CYC (CYCLOIDEA) in *Antirrhinum majus*, and the PCF1 and PCF2 (PROLIFERATING CELL FACTORS 1 and 2) in *Oryza sativa* [8,9]. Since its initial identification and characterization in 1999, the TCP family has become one of the focuses of plant studies due to its importance in the evolution and developmental control of plant form. The TCP domain contains a non-canonical basic-Helix-Loop-Helix (bHLH) structure involved in DNA binding, protein-protein interaction and protein nuclear localization [8,10]. The two amphipathic helical motifs are abundant in hydrophobic Ala, Leu and Trp residues while the disordered linking loop region contains acidic, polar and non-charged amino acids. By comparison, the most conserved basic region is rich in positively charged Lys and Arg amino acids [11]. The TCP transcription factor family can be further divided into two subfamilies, class I and class II, based mainly on amino acid sequence differences, especially in the basic region of the TCP domain [8]. According to the results of bioinformatics analysis, several class II TCP members also share an arginine-rich R domain outside the conserved TCP domain with unknown function, speculated to facilitate protein-protein interaction [8,12]. The DNA-Binding site selection assays revealed that the two TCP classes can specifically recognize and bind to slightly different but partly overlapping GC-rich DNA sequences which act as cis-element in a large number of plant genes. The DNA

binding sequences for class I is GGNCCCAC while class II prefer to bind the DNA motif G(T/C)GGNCCC [10,13-15]. To date, more than 20 TCP family members have been identified in a number of monocot and eudicot plants, such as *Arabidopsis*, *Oryza sativa*, *Vitis vinifera* and *Populus trichocarpa*.

TCP transcription factors play versatile functions in multiple biological processes during plant growth and development. It has been reported that many TCP transcription factors participate in the regulation of multiple aspects of plant development, such as gametophyte development [32-34], hormone signal transduction [29,35,36], mitochondrial biogenesis [37], regulation of the circadian clock [38,39], lateral branching [28,29,40], flower development[31,41,42], seed germination [43,44] and leaf development[14,31]. Class II TCP members have been found to function in a similar manner mainly by preventing plant growth and cell proliferation based on the mutation studies of multiple members in this subfamily [9,30,40,45-49], whereas the predicted role of class I members seems to promote plant growth and cell proliferation [10,50]. In *Arabidopsis*, mutation in *AtTCP18* (*BRC1*) gene led to a significant increase in the number of rosette branches while up-regulation of *AtTCP18* resulted in the inhibition of lateral branching, suggesting that *AtTCP18* plays a critical role in axillary bud outgrowth [29]. *AtTCP4* has been shown to influence early embryo development and recent evidence revealed that pollen grains produced by transgenic *Arabidopsis* line expressing hyper-activated *AtTCP4* genes cannot yield viable seeds, indicating that *AtTCP4* may regulate plant reproduction [32,33]. Functional analysis of *AtTCP1* showed that *AtTCP1* is involved in the regulation of Brassinosteroid hormone signaling pathway by positively controlling the expression of a key enzyme DWARF4 [35]. In a recent study, *AtTCP8* was proposed to be associated with mitochondrial biogenesis based on the evidence that *AtTCP8* is able to bind to the promoter region of *PNM1*, a gene encoding a newly identified pentatricopeptide repeat protein that function in the mitochondrial gene expression [37]. Yeast two-hybrid assays revealed the interaction between some TCP transcription factors (AtTCP2, AtTCP3, AtTCP11 and AtTCP15) and several regulatory components of the circadian clock, suggesting that TCP proteins may control or influence the circadian networks [38]. *AtTCP14* and *AtTCP15* were reported to regulate floral organ development and reduced expression of the two transcription factors resulted in phenotypic abnormalities in the three outer whorls and the gynoecia [31,41]. In addition, *AtTCP14* and *AtTCP15* were also found to regulate leaf development:

mutant *AtTCP14* and *AtTCP15* led to broader leaves towards the base and shorter petioles than the wild type [31].

Cotton comprises both diploid and tetraploid species, belonging to the *Gossypium* genus. The most commonly cultivated cotton species is upland cotton (*Gossypium hirsutum*), an AD tetraploid evolved from A-genome diploids such as *G. arboreum* and D-genome diploids like *G. raimondii* at around 1-2 million years ago. Widely cultivated in more than 100 countries, cotton is considered one of the most important fiber-producing and economic crops around the world, providing fiber for the textile industry and cooking oil extracted from its oil-rich seeds for food industry. The cotton industry is estimated to produce $133 billion in products and services annually, creating about 350 million jobs on farm or in the industry sectors. In spite of the economic and social importance of cotton and the critical role of TCP transcription factors in the control of plant cell proliferation and development, the research on cotton TCP family is much beyond on other plant species. Up to now, only two TCP family members have been functionally characterized in cotton, suggesting that TCP genes may play key roles in fiber development. In a recent study, Hao *et al* (2012) reported the functional characterization of a cotton TCP transcription factor GbTCP. According to this research, *GbTCP* was expressed in cotton fiber at much higher level than other tissues tested. Overexpression of *GbTCP* in *Arabidopsis* facilitated the initiation and elongation of root hair which has similar developmental mechanisms with cotton fiber, whereas RNAi silencing of *GbTCP* in cotton led to shorter fiber and low fiber quality, indicating that *GbTCP* played a significant role in fiber elongation [16]. In addition, Wang *et al.* demonstrated that the *GhTCP14* from upland cotton functions as a crucial regulator in auxin-mediated elongation of cotton fiber cells. However, no genome-wide characterization of TCP family members has been performed in cotton. The recent availability of the completed genome sequence of *Gossypium raimondii* [5] and *Gossypium arboreum*, two diploid cotton species, provides us with a great opportunity to identify and characterize TCP transcription factors in cotton genome.

Hypotheses

1. *G. raimondii* and *G. arboreum* may contain more TCP genes than *Arabidopsis.*

2. TCP genes within the same subfamily of the phylogenetic tree may share similar exon/intron organization and the motif structures may be highly conserved among the subfamilies.
3. TCP genes may be unevenly distributed across the chromosomes.
4. Segmental and tandem duplication may play a role in the evolution of TCP gene family in *G. raimondii* and *G. arboreum*.
5. TCP genes in *G. raimondii* and *G. arboreum* may be differentially expressed in different tissues.

Objectives

1. **Retrieve TCP transcription factor coding gene sequences and protein sequences from *G. raimondii* and *G. arboreum* genomes.**

The conserved TCP DNA-binding domain based on Hidden Markov Model (HMM) (PF03634) was obtained from Pfam protein family database (http://pfam.sanger.ac.uk/). In order to identify the TCP transcription factor coding genes of *G. raimondii* and *G. arboreum*, the HHM profile of TCP domain was subsequently employed as query to perform a HMMER search (http://hmmer.janelia.org/) against the *G. raimondii* genome and *G. arboreum* genome (E-value = 0.01). All redundant sequences were discarded from further analysis based on cluster W [17] alignment results, sequence identification numbers and chromosome location. Furthermore, to verify the reliability of the initial results, all non-redundant candidate TCP sequences were analyzed to confirm the presence of the conserved TCP domain using the InterproScan program [18]. The sequences of TCP family members in the genome of *Theobroma cacao*, *Vitis vinifera , Arabidopsis* and *Oryza sativa* were retrieved from PlantTFDB plant transcription factor database (http://planttfdb.cbi.pku.edu.cn/, v3.0).

2. **Perform a comprehensive analysis of the TCP transcription factor family in *G. raimondii* and *G. arboreum*, including phylogenetic relationships, chromosomal location, gene duplication status, gene structure and conserved motif.**

Phylogenetic analysis

Multiple sequence alignments were conducted on the amino acid sequences of TCP proteins in *G. raimondii*, *G. arboreum, Arabidopsis*, *Theobroma cacao*, *Vitis vinifera* and rice

genomes using Cluster X [19] with default settings. Subsequently, MEGA 6.0 software [20] was employed to construct an unrooted phylogenetic tree using the Neighbor-Joining (NJ) method with the following parameters: JTT model, pairwise gap deletion and 1,000 bootstraps. Furthermore, Maximum likelihood, Minimal Evolution and PhyML methods were also applied in the tree construction to validate the results from the NJ method. Additionally, a separate phylogenetic tree was constructed with all the TCP protein sequences in *G. raimondii* and *G. arboreum* for further analysis.

Analysis of chromosomal location and gene duplication

Information about the physical locations of all GrTCP genes on chromosomes were obtained through BLASTN searches against *G. raimondii* and *G. arboreum* genome database. All TCP genes were then mapped on the chromosome using the software MapInspect. The detection of TCP gene duplication events was also carried out. Paralogous TCP gene pairs in *G. raimondii* and *G. arboreum* were identified on the basis of alignment results. The criteria described in previous studies [21,22] were adopted: the shorter sequences covers over 70% of the longer sequence after alignment and the minimum identity of aligned regions is 70%. In addition, to further analyze gene duplication events, the synonymous substitution rate (Ks) and non-synonymous substitution rate (Ka) were calculated using the software DnaSp [23]. The date of duplication events was subsequently estimated according to the equation $T=Ks/2\lambda$. The approximate value for clock-like rate (λ) was 1.5 synonymous substitutions per 10^8 years for *G. arboreum* and *G. raimondii* [5].

Gene structure analysis and conserved motif identification

The TCP genomic sequences and CDS sequences was compared in gene structure display server program [24] to infer the exon/intron organization of TCP genes. TCP protein sequences in *G. raimondii* and *G. arboreum* were submitted to online Multiple Expectation maximization for Motif Elicitation (MEME) program [25] for identification of conserved protein motifs. The optimized MEME parameters were as follows: any number of repetitions, the optimum width from 6 to 250 and maximum number of motifs-20. The identified protein motifs were further annotated with ScanProsite [26].

3. **Investigate the expression profiles of cotton TCP genes in different tissues and at different developmental stages.**

To detect the expression profiles of TCP genes in *G. raimondii* and *G. arboreum,* total RNA was extracted from plant leaves, buds, shoots, sepals and fibers using the mirVanaTM miRNA Isolation Kit (Ambion, Austin, TX, USA), according to the manufacturer's instructions. The total RNA quantity and purity were assessed by a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). 1 μg of total RNA isolated from each tissue was reverse transcribed into cDNA using the TaqMan® MicroRNA Reverse Transcription Kit (Applied Biosystems, Foster City, CA, USA) and a poly-T primer. The real-time RT-PCR was then performed with a 7300 Real-Time PCR System (Applied Biosystems, Foster City, CA) according to the supplier's protocols. Each reaction mixture contains 2μL of DNase/RNase free water, 5 μL Real-Time SYBR Green PCR master mix, 1 μL diluted cDNA product from reverse transcription PCR reaction and 2 μL gene-specific primers. Three biological replicates were conducted for each tissue and each biological replicate was technically repeated three times. The thermal cycle applied was as follows: 95 °C for 10 min followed by 45 cycles of denature at 95 °C for 15 s and annealing and elongation at 60 °C for 60 s. The expression values of TCP genes tested were normalized with internal reference genes *TuA11* and *SAD1*. The relative expression levels (R) was calculated using the following equation: $R= 2^{-(Ct1 - Ct2)}$, where C_t1 stands for the C_t value of TCP genes while C_t2 is the C_t value of the reference gene. A heatmap for gene expression patterns was generated with the software MultiExperiment Viewer (MeV).

Relevance of the project

1. Benefits to scientific researchers

There is no comprehensive analysis of TCP transcription factor family in *G. raimondii* and *G. arboreum* up to now. The results of the present project will lay the foundation for functional characterization of TCP transcription family and uncover more knowledge related to the evolutionary relationships of TCP family among *G. raimondii*, *G. arboreum, Oryza sativa, Theobroma cacao*, *Vitis vinifera* and *Arabidopsis thaliana*. This will greatly benefit those researchers who study the regulation mechanisms of cotton growth and development. In addition, the genome-wide analysis of TCP gene family will also contribute to future studies on the identification and comprehensive analysis of the TCP transcription factor family in other species.

2. Benefits to society:

As a leading natural fiber source crop and one of the most important economic and oil crops, cotton has been widely cultivated in more than 70 developed and developing countries around the world. TCP transcription factor family plays an important role in cotton fiber growth and development. The present study can provide more potential TCP genes for genetic engineering, which will contribute a lot to improving cotton fiber yield and quality

References

1 Calkhoven, C. F. & Ab, G. Multiple steps in the regulation of transcription-factor level and activity. *Biochem. J.* **317 (Pt 2)**, 329-342 (1996).

2 Latchman, D. S. Transcription factors: an overview. *Int. J. Biochem. Cell Biol.* **29**, 1305-1312 (1997).

3 Schwechheimer, C., Zourelidou, M. & Bevan, M. W. Plant transcription factor studies. *Annu. Rev. Plant Phys.* **49**, 127-150 (1998).

4 Riechmann, J. L. *et al. Arabidopsis* transcription factors: Genome-wide comparative analysis among eukaryotes. *Science* **290**, 2105-2110 (2000).

5 Wang, K. *et al.* The draft genome of a diploid cotton Gossypium raimondii. *Nat. Genet.* **44**, 1098-1103, doi:10.1038/ng.2371 (2012).

6 Wray, G. A. *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**, 1377-1419, doi:10.1093/molbev/msg140 (2003).

7 Perez-Rodriguez, P. *et al.* PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res.* **38**, D822-827, doi:10.1093/nar/gkp805 (2010).

8 Cubas, P., Lauter, N., Doebley, J. & Coen, E. The TCP domain: a motif found in proteins regulating plant growth and development. *Plant J.* **18**, 215-222 (1999).

9 Martin-Trillo, M. & Cubas, P. TCP genes: a family snapshot ten years later. *Trends Plant Sci.* **15**, 31-39, doi:10.1016/j.tplants.2009.11.003 (2010).

10 Kosugi, S. & Ohashi, Y. PCF1 and PCF2 specifically bind to cis elements in the rice proliferating cell nuclear antigen gene. *Plant Cell* **9**, 1607-1619 (1997).

11 Yao, X., Ma, H., Wang, J. & Zhang, D. Genome-Wide Comparative Analysis and Expression Pattern of TCP Gene Families in *Arabidopsis* thaliana and Oryza sativa. *J. Integr. Plant Biol.* **49**, 885-897, doi:10.1111/j.1744-7909.2007.00509.x (2007).

12 Lupas, A., Vandyke, M. & Stock, J. Predicting Coiled Coils from Protein Sequences. *Science* **252**, 1162-1164 (1991).

13 Kosugi, S. & Ohashi, Y. DNA binding and dimerization specificity and potential targets for the TCP protein family. *Plant J.* **30**, 337-348 (2002).

14 Viola, I. L., Uberti Manassero, N. G., Ripoll, R. & Gonzalez, D. H. The *Arabidopsis* class I TCP transcription factor AtTCP11 is a developmental regulator with distinct DNA-binding properties due to the presence of a threonine residue at position 15 of the TCP domain. *Biochem. J.* **435**, 143-155, doi:10.1042/BJ20101019 (2011).

15 Schommer, C. *et al.* Control of jasmonate biosynthesis and senescence by miR319 targets. *PLoS Biol.* **6**, e230, doi:10.1371/journal.pbio.0060230 (2008).

16 Hao, J. *et al.* GbTCP, a cotton TCP transcription factor, confers fibre elongation and root hair development by a complex regulating system. *J. Exp. Bot.* **63**, 6267-6281, doi:10.1093/jxb/ers278 (2012).

17 Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680 (1994).

18 Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116-120, doi:10.1093/nar/gki442 (2005).

19 Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876-4882 (1997).
20 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725-2729, doi:10.1093/molbev/mst197 (2013).
21 Yang, S., Zhang, X., Yue, J. X., Tian, D. & Chen, J. Q. Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genomics.* **280**, 187-198, doi:10.1007/s00438-008-0355-0 (2008).
22 Gu, Z., Cavalcanti, A., Chen, F. C., Bouman, P. & Li, W. H. Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Mol. Biol. Evol.* **19**, 256-262 (2002).
23 Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452, doi:10.1093/bioinformatics/btp187 (2009).
24 Guo, A. Y., Zhu, Q. H., Chen, X. & Luo, J. C. [GSDS: a gene structure display server]. *Yi Chuan* **29**, 1023-1026 (2007).
25 Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369-373, doi:10.1093/nar/gkl198 (2006).
26 de Castro, E. *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* **34**, W362-365, doi:10.1093/nar/gkl124 (2006).
27 Aguilar-Martinez, J. A. & Sinha, N. Analysis of the role of *Arabidopsis* class I TCP genes AtTCP7, AtTCP8, AtTCP22, and AtTCP23 in leaf development. *Front. Plant Sci.* **4**, 406, doi:10.3389/fpls.2013.00406 (2013).
28 Takeda, T. *et al.* The OsTB1 gene negatively regulates lateral branching in rice. *Plant J.* **33**, 513-520 (2003).
29 Aguilar-Martinez, J. A., Poza-Carrion, C. & Cubas, P. *Arabidopsis* BRANCHED1 acts as an integrator of branching signals within axillary buds. *Plant Cell* **19**, 458-472, doi:10.1105/tpc.106.048934 (2007).
30 Palatnik, J. F. *et al.* Control of leaf morphogenesis by microRNAs. *Nature* **425**, 257-263, doi:10.1038/nature01958 (2003).
31 Kieffer, M., Master, V., Waites, R. & Davies, B. TCP14 and TCP15 affect internode length and leaf shape in *Arabidopsis*. *Plant J.* **68**, 147-158, doi:10.1111/j.1365-313X.2011.04674.x (2011).
32 Pagnussat, G. C. *et al.* Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. *Development* **132**, 603-614 (2005).
33 Sarvepalli, K. & Nath, U. Hyper-activation of the TCP4 transcription factor in *Arabidopsis* thaliana accelerates multiple aspects of plant maturation. *Plant J.* **67**, 595-607 (2011).
34 Takeda, T. *et al.* RNA interference of the *Arabidopsis* putative transcription factor TCP16 gene results in abortion of early pollen development. *Plant Mol. Biol.* **61**, 165-177 (2006).

35 Guo, Z. X. *et al.* TCP1 Modulates Brassinosteroid Biosynthesis by Regulating the Expression of the Key Biosynthetic Gene DWARF4 in *Arabidopsis* thaliana. *Plant Cell* **22**, 1161-1173 (2010).
36 Yanai, O., Shani, E., Russ, D. & Ori, N. Gibberellin partly mediates LANCEOLATE activity in tomato. *Plant J.* **68**, 571-582 (2011).
37 Hammani, K. *et al.* An *Arabidopsis* Dual-Localized Pentatricopeptide Repeat Protein Interacts with Nuclear Proteins Involved in Gene Expression Regulation. *Plant Cell* **23**, 730-740 (2011).
38 Giraud, E. *et al.* TCP Transcription Factors Link the Regulation of Genes Encoding Mitochondrial Proteins with the Circadian Clock in *Arabidopsis* thaliana. *Plant Cell* **22**, 3921-3934 (2010).
39 Pruneda-Paz, J. L., Breton, G., Para, A. & Kay, S. A. A Functional Genomics Approach Reveals CHE as a Component of the *Arabidopsis* Circadian Clock. *Science* **323**, 1481-1485 (2009).
40 Hubbard, L., McSteen, P., Doebley, J. & Hake, S. Expression patterns and mutant phenotype of teosinte branched1 correlate with growth suppression in maize and teosinte. *Genetics* **162**, 1927-1935 (2002).
41 Uberti-Manassero, N. G., Lucero, L. E., Viola, I. L., Vegetti, A. C. & Gonzalez, D. H. The class I protein AtTCP15 modulates plant development through a pathway that overlaps with the one affected by CIN-like TCP proteins. *J. Exp. Bot.* **63**, 809-823, doi:10.1093/jxb/err305 (2012).
42 Koyama, T., Ohme-Takagi, M. & Sato, F. Generation of serrated and wavy petals by inhibition of the activity of TCP transcription factors in *Arabidopsis* thaliana. *Plant Signal. Behav.* **6**, 697-699 (2011).
43 Tatematsu, K., Nakabayashi, K., Kamiya, Y. & Nambara, E. Transcription factor AtTCP14 regulates embryonic growth potential during seed germination in *Arabidopsis* thaliana. *Plant J.* **53**, 42-52, doi:10.1111/j.1365-313X.2007.03308.x (2008).
44 Rueda-Romero, P., Barrero-Sicilia, C., Gomez-Cadenas, A., Carbonero, P. & Onate-Sanchez, L. *Arabidopsis* thaliana DOF6 negatively affects germination in non-after-ripened seeds and interacts with TCP14. *J. Exp. Bot.* **63**, 1937-1949, doi:10.1093/jxb/err388 (2012).
45 Doebley, J., Stec, A. & Hubbard, L. The evolution of apical dominance in maize. *Nature* **386**, 485-488 (1997).
46 Luo, D., Carpenter, R., Vincent, C., Copsey, L. & Coen, E. Origin of floral asymmetry in Antirrhinum. *Nature* **383**, 794-799 (1996).
47 Crawford, B. C. W., Nath, U., Carpenter, R. & Coen, E. S. CINCINNATA controls both cell differentiation and growth in petal lobes and leaves of antirrhinum. *Plant Physiol.* **135**, 244-253 (2004).
48 Nath, U., Crawford, B. C. W., Carpenter, R. & Coen, E. Genetic control of surface curvature. *Science* **299**, 1404-1407 (2003).
49 Lewis, J. M. *et al.* Overexpression of the maize Teosinte Branched1 gene in wheat suppresses tiller development. *Plant Cell Rep.* **27**, 1217-1225 (2008).
50 Li, C. X., Potuschak, T., Colon-Carmona, A., Gutierrez, R. A. & Doerner, P. *Arabidopsis* TCP20 links regulation of growth and cell division control pathways. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12978-12983 (2005).

51 *Arabidopsis* Genome, I. Analysis of the genome sequence of the flowering plant *Arabidopsis* thaliana. *Nature* **408**, 796-815, doi:10.1038/35048692 (2000).
52 Zhang, J. Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**, 292-298, doi:10.1016/s0169-5347(03)00033-8 (2003).
53 Flagel, L. E. & Wendel, J. F. Gene duplication and evolutionary novelty in plants. *New Phytol.* **183**, 557-564, doi:10.1111/j.1469-8137.2009.02923.x (2009).
54 Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant cell* **16**, 1667-1678, doi:10.1105/tpc.021345 (2004).
55 Prince, V. E. & Pickett, F. B. Splitting pairs: the diverging fates of duplicated genes. *Nat. Rev. Genet.* **3**, 827-837, doi:10.1038/nrg928 (2002).
56 Wang, M. Y. *et al.* The cotton transcription factor TCP14 functions in auxin-mediated epidermal cell differentiation and elongation. *Plant Physiol.* **162**, 1669-1680, doi:10.1104/pp.113.215673 (2013).

CHAPTER 2: Genome-Wide Identification and Expression Analysis of TCP Transcription Factors in *Gossypium Raimondii*

Abstract

Plant-specific TEOSINTE-BRANCHED1/CYCLOIDEA/PCF (TCP) transcription factor family plays versatile functions in multiple aspects of plant growth and development. However, no systematical study of this family has been conducted in cotton. In this study, we performed for the first time the genome-wide identification and expression analysis of the TCP transcription factor family in *Gossypium raimondii*. A total of 38 non-redundant cotton TCP encoding genes were identified. The TCP transcription factors were divided into eleven subgroups based on phylogenetic analysis. Most TCP genes within the same subfamily demonstrated similar exon and intron organization and the motif structures were highly conserved among the subfamilies. Additionally, the chromosomal distribution pattern revealed that TCP genes were unevenly distributed across 11 out of the 13 chromosomes; segmental duplication is a predominant duplication event for TCP genes and the major contributor to the expansion of TCP gene family in *G. raimondii*. Moreover, the expression profiles of TCP genes shed light on their functional divergence.

Keywords: Cotton, *Gossypium raimondii*, TCP, Expression profiles

Introduction

Regulation of gene transcription, one of the most complex activities in cells, plays a significant role in a wide variety of biological processes, such as cell growth, cell cycle control, signal transduction, metabolic and physiological balance, and response to environmental stimuli [1-5]. Among many mechanisms of transcriptional regulation of gene expression, transcription factors are considered to be the most important. Transcription factors are a diverse family of regulatory proteins with specific DNA-binding domains involved in the regulation of many cellular processes by either stimulating or repressing transcription of the related genes [2,3,6]. Up to now, more than 60 transcription factor families have been identified in plants [7].

The TCP proteins are a family of transcription factors exclusive to higher plants and involved in the regulation of cell growth and proliferation [8,9]. This class of transcription factors are featured by a highly conserved 59-residue-long DNA-binding motif at the N-terminus called TCP domain, which is named after four founding members: TB1 (TEOSINTE BRANCHED 1) in *Zea mays*, CYC (CYCLOIDEA) in *Antirrhinum majus*, and the PCF1 and PCF2 (PROLIFERATING CELL FACTORS 1 and 2) in *Oryza sativa* [8,9]. The TCP domain contains a non-canonical basic-Helix-Loop-Helix (bHLH) structure involved in DNA binding, protein-protein interaction and protein nuclear localization [8,10]. The two amphipathic helical motifs are abundant in hydrophobic Ala, Leu and Trp residues while the disordered linking loop region contains acidic, polar and non-charged amino acids. By comparison, the most conserved basic region is rich in positively charged Lys and Arg amino acids [11]. The TCP transcription factor family can be further divided into two subfamilies, class I and class II, based mainly on amino acid sequence differences, especially in the basic region of the TCP domain [8]. According to the results of bioinformatics analysis, several class II TCP members also share an arginine-rich R domain outside the conserved TCP domain with unknown function, speculated to facilitate protein-protein interaction [8,12]. The DNA-Binding site selection assays revealed that the two TCP classes can specifically recognize and bind to slightly different but partly overlapping GC-rich DNA sequences which act as cis-element in a large number of plant genes. The DNA binding sequences for class I is GGNCCCAC while class II prefer to bind the DNA motif G(T/C)GGNCCC [10,13-15].

Widely cultivated in more than 100 countries, cotton is considered one of the most important fiber-producing and economic crops around the world, providing fiber for the textile industry and cooking oil extracted from its oil-rich seeds for food industry. The cotton industry is estimated to produce $133 billion in products and services annually, creating about 350 million jobs on farm or in the industry sectors. In spite of the economic and social importance of cotton and the critical role of TCP transcription factors in the control of plant cell proliferation and development, the research on cotton TCP family is much beyond on other plant species. In a recent study, Hao *et al* (2012) reported the functional characterization of a cotton TCP transcription factor GbTCP. According to this research, *GbTCP* was expressed in cotton fiber at much higher level than other tissues tested. Overexpression of *GbTCP* in *Arabidopsis* facilitated the initiation and elongation of root hair which has similar developmental mechanisms with cotton fiber, whereas RNAi silencing of *GbTCP* in cotton led to shorter fiber and low fiber quality, indicating that *GbTCP* played a significant role in fiber elongation [16]. Up to now, however, no genome-wide characterization of TCP family members has been performed in cotton. The recent availability of the completed genome sequence of *Gossypium raimondii* [5], a diploid cotton species, provides us with a great opportunity to identify and characterize TCP transcription factors in cotton genome.

In the present study, we performed for the first time the comprehensive analysis of the TCP transcription factor family in *G. raimondii*. A total of 38 non-redundant TCP transcription factor encoding genes were identified in the genome of *G. raimondii* and were subsequently subjected to a systematic analysis, including phylogenetic relationships, chromosomal location, gene duplication status, gene structure, conserved motif and expression profiling. On the basis of the expression profiles of TCP members in *G. raimondii* and the phylogenetic analysis among the TCP domain proteins in *Arabidopsis*, rice and *G. raimondii*, the functions of GrTCPs were predicted. Besides, it is also remarkable that the expansion of TCP family in *G. raimondii* may be caused mainly by segmental duplication and is not associated with tandem duplication. In a word, our genome-wide analysis of TCP gene family will contribute to future studies on the functional characterization of TCP proteins in *G. raimondii* as well as the identification and comprehensive analysis of the TCP transcription factor family in other species.

Materials and Methods

Sequence retrieval for TCP proteins

The conserved TCP DNA-binding domain based on Hidden Markov Model (HMM) (PF03634) was obtained from Pfam protein family database (http://pfam.sanger.ac.uk/). In order to identify the TCP transcription factor coding genes of *G. raimondii*, the HHM profile of TCP domain was subsequently employed as query to perform a HMMER search (http://hmmer.janelia.org/) against the *G. raimondii* genome derived from Phytozome (http://www.phytozome.net/) (E-value = 0.01). All redundant sequences were discarded from further analysis based on cluster W [17] alignment results, sequence identification numbers and chromosome location. Furthermore, to verify the reliability of the initial results, all non-redundant candidate TCP sequences were analyzed to confirm the presence of the conserved TCP domain using the InterproScan program [18]. The sequences of TCP family members in the genome of *Arabidopsis* and *Oryza sativa* were retrieved from PlantTFDB plant transcription factor database (http://planttfdb.cbi.pku.edu.cn/, v3.0).

Phylogenetic analysis

Multiple sequence alignments were conducted on the amino acid sequences of TCP proteins in *G. raimondii*, *Arabidopsis* and rice genomes using Cluster X [19] with default settings. Subsequently, MEGA 6.0 software [20] was employed to construct an unrooted phylogenetic tree based on alignments using the Neighbor-Joining (NJ) method with the following parameters: JTT model, pairwise gap deletion and 1,000 bootstraps. Furthermore, Maximum likelihood, Minimal Evolution and PhyML methods were also applied in the tree construction to validate the results from the NJ method. Additionally, a separate phylogenetic tree was constructed with all the TCP protein sequences in *G. raimondii* for further analysis.

Analysis of chromosomal location and gene duplication

Information about the physical locations of all GrTCP genes on chromosomes were obtained through BLASTN searches against *G. raimondii* genome database in Phytozome (http://www.phytozome.net/cotton.php). All GrTCP genes were then mapped on the chromosome using the software MapInspect. The detection of TCP gene duplication events was also carried out. Paralogous TCP gene pairs in *G. raimondii* were identified on the basis of alignment results. The criteria described in previous studies [21,22] were adopted: the shorter sequences covers over 70% of the longer sequence after alignment and the minimum identity of aligned regions is 70%. In addition, to further analyze gene duplication events, the synonymous

substitution rate (Ks) and non-synonymous substitution rate (Ka) were calculated using the software DnaSp [23]. The date of duplication events was subsequently estimated according to the equation T=Ks/2λ. The approximate value for clock-like rate (λ) was 1.5 synonymous substitutions per 10^8 years for *G. raimondii* [5].

Gene structure analysis and conserved motif identification

The TCP genomic sequences and CDS sequences extracted from Phytozome was compared in gene structure display server program [24] to infer the exon/intron organization of TCP genes. TCP protein sequences in *G. raimondii* were submitted to online Multiple Expectation maximization for Motif Elicitation (MEME) program [25] for identification of conserved protein motifs. The optimized MEME parameters were as follows: any number of repetitions, the optimum width from 6 to 250 and maximum number of motifs-20. The identified protein motifs were further annotated with ScanProsite [26].

RNA isolation and Real-time quantitative RT-PCR analysis

To detect the expression profiles of TCP genes in *G. raimondii*, total RNA was extracted from plant leaves, buds, shoots and sepals using the mirVanaTM miRNA Isolation Kit (Ambion, Austin, TX, USA), according to the manufacturer's instructions. The total RNA quantity and purity were assessed by a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). 1 μg of total RNA isolated from each tissue was reverse transcribed into cDNA using the TaqMan® MicroRNA Reverse Transcription Kit (Applied Biosystems, Foster City, CA, USA) and a poly-T primer. The real-time RT-PCR was then performed with a 7300 Real-Time PCR System (Applied Biosystems, Foster City, CA) according to the supplier's protocols. Each reaction mixture contains 2μL of DNase/RNase free water, 5 μL Real-Time SYBR Green PCR master mix, 1 μL diluted cDNA product from reverse transcription PCR reaction and 2 μL gene-specific primers. Three biological replicates were conducted for each tissue and each biological replicate was technically repeated three times. The thermal cycle applied was as follows: 95 °C for 10 min followed by 45 cycles of denature at 95 °C for 15 s and annealing and elongation at 60 °C for 60 s. The expression values of TCP genes tested were normalized with an internal reference gene *TuA11*. The relative expression levels (R) was calculated using the following equation: $R= 2^{-(Ct1 - Ct2)}$, where C_t1 stands for the C_t value of TCP genes while C_t2 is the C_t value of the reference gene. A heatmap for gene expression patterns was generated with the software MultiExperiment Viewer (MeV).

Results

Identification of TCP genes

In order to identify the TCP transcription factor coding genes of *G. raimondii*, the HMM profile of TCP domain (PF03634) was employed as query to perform a hmmer search against the *G. raimondii* genome (http://www.phytozome.net/cotton). Originally, 62 candidate TCP genes were identified in *G. raimondii*. Among them, 24 redundant sequences were discarded from further analysis based on their sequence similarity. Subsequently, with the aim to verify the reliability of the initial results, a survey was conducted to confirm the existence of the conserved TCP domain with InterproScan [18]. The results showed that all of the 38 putative TCP genes contained conserved TCP domain. Due to the lack of standard annotation designated to the 38 TCP genes in the *G. raimondii*, we named them *GrTCP1* to *GrTCP25* according to the *Arabidopsis* TCP proteins with highest sequence similarity and following the nomenclature system applied to *Arabidopsis*. The length of the 38 newly identified TCP transcription factors varied from 196 to 549 amino acids with an average of 353.5 amino acids. Other characteristics of TCP transcription factors in *G. raimondii*, including isoelectric point (pI), molecular weight (Mw) and chromosome location, were listed in Table 1.1.

Phylogenetic analysis

To get a better understanding of the evolutionary history and phylogenetic relationships of TCP transcription factor family in *G. raimondii*, an unrooted phylogenetic tree was constructed with Neighbor-Joining method on the basis of multiple sequence alignment of 38 *G. raimondii* TCP protein sequences with all TCP sequences from *Arabidopsis* and rice, including 24 *Arabidopsis* TCP protein sequences and 22 rice TCP protein sequences (Figure 1.1). The bootstrap values for some nodes of the NJ tree were low as a result of relatively large number of sequences, which was also shown in previous reports [9,11]. Therefore, we sought other evidence to verify the reliability of our phylogenetic tree. The phylogenetic trees of TCP transcription family were reconstructed with Maximum likelihood, Minimal Evolution and PhyML methods. The trees produced by the three methods mentioned above were almost identical with only minor differences at some branches, suggesting that the four methods were largely consistent with each other. Besides, the analysis of gene structure, conserved motif structure and expression profiles

were also used to confirm the validity of the phylogenetic tree. Considering the great similarity among these tree topologies as well as previous studies [9,11], the NJ tree was employed for further analysis.

According to the NJ phylogenetic tree (Figure 1.1), the TCP transcription factor family was divided into eleven subgroups designated Group A to Group K. According to their sequence features within and outside the TCP domain, GrTCPs in Group A, B, C, D, E, F, and G belong to Class I subfamily while GrTCPs in the rest groups belong to Class II subfamily[8]. Group A, the largest clade among all subgroups, contained 12 members, representing 14.3 % of the total TCPs; Group E constituted the smallest clade, containing 3 members. Generally speaking, the TCP family members showed an interspersed distribution in most clades, indicating that the TCP family expanded before the divergence of the lineages. Additionally, the TCPs were not evenly distributed in some clades in *G. raimondii*, *Arabidopsis* and rice. Many *Arabidopsis* TCPs had two or more counterparts in *G. raimondii*, suggesting that GrTCP genes duplicated after the divergence of *G. raimondii* and *Arabidopsis*. For example, Group A contained seven *G. raimondii* TCPs but there were only three *Arabidopsis* members; Group D contained four *G. raimondii* TCPs but there were only two *Arabidopsis* TCPs. Specifically, Group F contained five *G. raimondii* TCPs but there was only one *Arabidopsis* TCP and no TCP was identified in rice genome, which implied that this group was either acquired after the divergence of monocots and dicots or lost in rice. By comparison, the rice TCPs were overrepresented in Group K, which contained six rice TCPs, while there were only two *G. raimondii* members and two *Arabidopsis* members in the same group. Some groups had almost equal number of TCP genes in the three species, such as Group B, C, and E (Figure 1.1).

Many *Arabidopsis* TCPs with similar functions tended to cluster into the same clade, which may imply that TCP family members within the same clade had similar functions in *G. raimondii*. For example, all AtTCPs in Group A and Group B (AtTCP8, AtTCP14, AtTCP15, AtTCP22, AtTCP23), which clustered with ten GrTCPs (Figure 1.1), play an important role in the regulation of leaf development by modulating gene networks involved in cell-cycle control and shoot apical meristem (SAM) maintenance [27]. AtTCPs in Group H function in the process of lateral branching which determines shoot architecture [28,29]. All AtTCPs in Group I and Group K (AtTCP2, AtTCP3, AtTCP4, AtTCP10, AtTCP24), which cluster with five GrTCPs (Figure 1.1),

are down regulated by miRNA319 and act as negative cell proliferation factors in the regulation of leaf margins [30].

Chromosomal location and gene duplication

To determine the chromosomal distribution of the TCP genes in *G. raimondii*, the physical locations of all GrTCP genes on chromosomes were obtained through BLASTN searches against *G. raimondii* genome database in Phytozome (http://www.phytozome.net/cotton.php). Among the 38 GrTCP genes, a total of 36 genes were distributed across 11 out of the 13 *G. raimondii* chromosomes, while the rest two (*GrTCP15b* and *GrTCP16*) were anchored on unmapped scaffolds (Figure 1.2). Generally speaking, the number of GrTCP genes on each chromosome appeared to be uneven, ranging widely from 0 to 8 genes per chromosome. For example, chromosome 8 contained the highest number of 8 GrTCPs, accounting for 21.1% of the total GrTCP genes, followed by 4 GrTCPs on each of chromosome 1, 6, 9 and 13, whereas relatively low number of GrTCP genes were found on several chromosomes, including 2 genes on each of chromosome 2 and 11, and 1 genes on each of chromosome 4 and 5. By contrast, GrTCP genes were not observed on chromosome 3 and 10 (Figure 1.2).

Given the importance of gene duplication in the amplification of gene families, potential duplication events involved in the evolution of *G. raimondii* genome were analyzed to shed light on the mechanism behind the expansion of the GrTCP gene family. On the basis of protein sequence identities, 19 pairs of putative paralogous GrTCP genes were identified, accounting for more than 70% of the entire GrTCP gene family and thereby supporting the hypothesis that putative gene duplication events are the main causes of the expansion of the GrTCP gene family. These gene pairs are in the same clade of the phylogenetic tree with high degree of protein sequence identities. For instance, the sequence of GrTCP15a covers 100% of that of GrTCP15b after alignment and the identity of aligned region is 98%, while the protein sequence identity of GrTCP7a and GrTCP7b is 88%. Among these paralogous gene pairs, 18 pairs are located on different chromosomes, suggesting a high number of segmental duplication events, whereas no traceable duplication events could be determined for another gene pair because one gene of this pair was anchored on unmapped scaffolds (Figure 1.2). Interestingly, six genes (*GrTCP3*, *GrTCP6*, *GrTCP7a*, *GrTCP7b*, *GrTCP7c* and *GrTCP20c*) participated in two segmental

duplication events (e.g. *GrTCP20a* / *GrTCP20b*/ *GrTCP20c*, *GrTCP7a*/ *GrTCP7b*/ *GrTCP7c* and *GrTCP3*/ *GrTCP4*/ *GrTCP10*). In contrast, no tandem duplication events were observed in these duplicated pairs (Figure 1.2).

In the present study, we further calculated the approximate dates of duplication events with the DnaSP program. The results showed that segmental duplications of GrTCP genes occurred between 11.28 Mya (million years ago) to 36.51 Mya, with an average of 19.83 Mya.

Gene structure and conserved motifs

With the aim to gain further insights into the evolutionary relationships among GrTCP genes, we investigated the exon/intron structures of individual GrTCP genes by alignment of cDNA sequences and corresponding genomic DNA sequences. As illustrated in Figure 1.3b, 32 out of 38 GrTCP genes had no intron, while the other GrTCP genes possess one intron, with the exception of *GrTCP25* containing five introns. Additionally, an unrooted phylogenetic tree was constructed with GrTCP protein sequences to determine if the exon/intron organization of GrTCP genes is consistent with the phylogenetic subfamilies (Figure 1.3a). As expected, most GrTCP genes within the same subfamily demonstrated very similar exon/intron distribution patterns in terms of exon length and intron number. For example, most GrTCP gene in subfamily A, B, C, D, E and K had only one exon of similar length without intron, whereas members within subfamily H contain one intron, except for *GrTCP12*, which possesses no intron. By comparison, GrTCP genes in subfamily G showed great variability in exon length and intron number (Figure 1.3a and 1.3b).

We further searched for the conserved motifs in GrTCP proteins by MEME program to obtain more insights into the diversity of motif compositions among GrTCPs. As shown in Figure 3c, a total of 20 conserved motifs designated as motif 1 to motif 20 were identified. Most of GrTCP proteins within the same subfamily shared similar motif compositions while high divergence was observed among different subfamilies, implying that the GrTCP members within the same subfamily may perform similar functions and that some motifs may plan an important role in the subfamily-specific- functions. For example, all GrTCPs in subfamily C possess motif 1, 2, 4, 8, 12 and 15 while all members in subfamily I contain motif 1, 2, 9, 14, 17 and 18 (Figure 1.3c). In addition, some motifs were exclusively present in a particular subfamily, suggesting that these motifs may contribute to the specific function of that subfamily. For instance, motif 20 for subfamily A, motif 6 for subfamily F and motif 14, 17 and 18 for subfamily I (Figure 1.3c).

Moreover, the program ScanProsite was employed to annotate the identified 20 motifs. However, few motifs hit for PROSITE (release 20.103) motif database. Therefore, the functions of most motifs are still left unknown. The only motif that matched to protein sequences in the ScanProsite database was motif 1, which was annotated as the conserved TCP domain and was uniformly observed in all GrTCP proteins. Generally speaking, the consistency of the motif compositions of GrTCP proteins as well as the exon/intron structures of most GrTCP genes with the phylogenetic subfamilies further supported the close evolutionary relationships among GrTCPs as well as the reliability of our phylogenetic analysis.

Expression profiles of TCP genes in G. raimondii

To investigate the tissue-specific expression profiles of TCP genes in *G. raimondii*, the quantitative real time PCR (qRT-PCR) was performed for different organs, including leaf, flower bud, shoot and sepal. As indicated in Figure 1.4 and Figure 1.5, some GrTCP genes were differentially expressed in the four tissues tested while other GrTCP genes showed similar expression patterns in different tissues, which may indicate functional divergence of GrTCP genes during plant development. For example, *GrTCP2*, *GrTCP3*, *GrTCP13b*, *GrTCP15c*, *GrTCP19a* and *GrTCP23* were constitutively expressed in every tissue tested at very high level, implying that these genes may play regulatory roles at multiple development stages, whereas *GrTCP6*, *GrTCP9a*, *GrTCP20a* and *GrTCP24* were expressed at very low level in all tissues examined, which suggested that they may be primarily expressed in other organs not tested or under special conditions (Figure 1.4 and Figure 1.5). In contrast, the expression levels of *GrTCP20a*, *GrTCP20b* and *GrTCP20c* were very high in leaf and bud and were relatively low in shoot and sepal, indicating that they may play an important role in the development of leaf and bud. A similar expression profile was also found for *GrTCP1* and *GrTCP25*. In addition, some genes were exclusively highly expressed in a specific tissue. For example, *GrTCP10*, *GrTCP12*, *GrTCP14b*, *GrTCP18b* and *GrTCP20d* were relatively highly expressed in leaf while *GrTCP7a*, *GrTCP7c*, *GrTCP9b* and *GrTCP14* were exclusively expressed in shoot at very high level, implying their specific roles in the corresponding tissues (Figure 1.4 and Figure 1.5). In general, the GrTCP genes that are highly expressed in specific tissues may be involved in the regulation of plant development. For instance, *GrTCP15c* were relatively highly expressed in leaf and bud, suggesting that it may play a role in the development of the two tissues. According to previous studies, *AtTCP15*, the homologous counterparts of *GrTCP15c*, is involved in the regulation of

leaf development [31], which supported our hypothesis. However, further studies are still needed to unravel the divergent roles of GrTCP genes.

Discussion

TCP transcription factors play important roles in plants

TCP transcription factors are a class of plant-specific transcription factors, which play versatile functions in multiple biological processes during plant growth and development (Figure 1.6). It has been reported that many TCP transcription factors participate in the regulation of multiple aspects of plant development, such as gametophyte development [32-34], hormone signal transduction [29,35,36], mitochondrial biogenesis [37], regulation of the circadian clock [38,39], lateral branching [28,29,40], flower development[31,41,42], seed germination [43,44] and leaf development[14,31]. Class II TCP members have been found to function in a similar manner mainly by preventing plant growth and cell proliferation based on the mutation studies of multiple members in this subfamily [9,30,40,45-49], whereas the predicted role of class I members seems to promote plant growth and cell proliferation [10,50]. In *Arabidopsis*, mutation in *AtTCP18* (*BRC1*) gene led to a significant increase in the number of rosette branches while up-regulation of *AtTCP18* resulted in the inhibition of lateral branching, suggesting that *AtTCP18* plays a critical role in axillary bud outgrowth [29]. *AtTCP4* has been shown to influence early embryo development and recent evidence revealed that pollen grains produced by transgenic *Arabidopsis* line expressing hyper-activated *AtTCP4* genes cannot yield viable seeds, indicating that *AtTCP4* may regulate plant reproduction [32,33]. Functional analysis of *AtTCP1* showed that *AtTCP1* is involved in the regulation of Brassinosteroid hormone signaling pathway by positively controlling the expression of a key enzyme DWARF4 [35]. In a recent study, *AtTCP8* was proposed to be associated with mitochondrial biogenesis based on the evidence that *AtTCP8* is able to bind to the promoter region of *PNM1*, a gene encoding a newly identified pentatricopeptide repeat protein that function in the mitochondrial gene expression [37]. Yeast two-hybrid assays revealed the interaction between some TCP transcription factors (AtTCP2, AtTCP3, AtTCP11 and AtTCP15) and several regulatory components of the circadian clock, suggesting that TCP proteins may control or influence the circadian networks [38]. *AtTCP14* and *AtTCP15* were reported to regulate floral organ development and reduced expression of the two transcription factors resulted in phenotypic abnormalities in the three outer whorls and the gynoecia [31,41]. In

addition, *AtTCP14* and *AtTCP15* were also found to regulate leaf development: mutant *AtTCP14* and *AtTCP15* leads to broader leaves towards the base and shorter petioles than the wild type [31].

TCP transcription factors were widely existed in cotton

In the present study, a total of 38 TCP genes were identified from *G. raimondii*; the number of TCP genes in *G. raimondii* was higher than that in *Arabidopsis* (24) and in rice (22) [9]. The number of TCP genes in *G. raimondii* is approximately 1.58 times that in *Arabidopsis*, which is in strong agreement with the fact that the protein coding genes in *G. raimondii* genome (40,976 genes) is about 1.6 times that in *Arabidopsis* (25,498 genes) [5,51]. It is found that many TCP genes in *Arabidopsis* have two or more counterparts in *G. raimondii*, suggesting that the expansion of TCP family in *G. raimondii* may be caused by gene duplication events such as segmental duplication, tandem duplication and transposition events.

Gene duplication, an outstanding feature of genomic architecture, plays a significant role in the process of plant genomic and organismal evolution, generating raw genetic material necessary for mutation, genetic drift and selection, and contributing to the origin of new gene functions and the evolution of gene networks [52,53]. It has been demonstrated that the expansion of gene families is mainly attributed to gene duplication events on various scales, including tandem duplication, segmental duplication, transposition events and whole-genome duplication [52,53]. Our results indicate that segmental duplication is a predominant duplication event for TCP genes and the major contributor to the expansion of TCP gene family in *G. raimondii*. It was reported that *G. raimondii* genome has undergone at least two rounds of genome-wide duplication, an ancient paleohexaploidization event at approximately 130.8 Mya and a recent whole-genome duplication event at around 13.3-20.0 Mya [5]. The average duplication dates of GrTCP genes is very close to the recent whole-genome duplication date of *G. raimondii*, suggesting that large-scale genome duplication events may also contribute to the expansion of GrTCP family. In addition, according to a recent study, the split of *G. raimondii*/ *Arabidopsis* and *G. raimondii*/ *T. cacao* occurred at approximately 82.3 million years ago and 33.7 million years ago, respectively [5]. Since the duplication of GrTCPs originated from 11.28 to 36.51 million years ago, most of the GrTCP genes duplicated after the divergence of *G. raimondii*/ *Arabidopsis* and *G. raimondii*/ *T. cacao*. Summarizedly, our results indicate that both segmental duplication and whole-genome duplication contribute to the expandedness of TCP family in *G. raimondii*.

Evolutionary conservation and divergence of the TCP family in G. raimondii

According to recent studies, *G. raimondii* and *Arabidopsis* diverged from a common ancestor at approximately 82.3 million years ago, which is followed by paleopolyploidy events in both species, contributing to evolutionary innovation[5,54]. Our studies showed that many TCP genes in *Arabidopsis* have two or more counterparts in *G. raimondii* with high protein sequence similarity, implying that the TCP genes may undergo differential expansion in *G. raimondii* and *Arabidopsis*. Through comparison of the exon/intron organization of individual TCP family members in the two species, we observed that their gene structures exhibited high similarities (Figure 1.3b and Supplementary Fig. S1.2). Among ten pairs of TCP genes with high protein sequence identity and query coverage, seven pairs exhibited conserved gene structure in terms of exon length and intron numbers. While most pairs showed similar exon/intron structure, a few displayed some degree of divergence. For example, *GrTCP12* contained one exon without intron, whereas its counterpart in *Arabidopsis*, *AtTCP12*, possessed one exon and one intron. Such variation may be caused by single intron loss or gain during the process of structural evolution. In addition, we further analyzed the gene structure and conserved motifs of paralogous pairs of TCPs in *G. raimondii* to shed light on the diversification of TCP genes. Of the 19 paralogous pairs, 14 pairs of GrTCPs shared conserved exon/intron organization (Figure 1.3b). Similar to the gene structure, most paralogous pairs of GrTCPs also exhibited conserved motif composition, with only several unique motifs observed in some GrTCP members, such as motif 20 for GrTCP15a, motif 3 for GrTCP9b and motif 9 for GrTCP18b and GrTCP14a (Figure 1.3c). These specific motifs may contribute to the neofunctionalization (in which one paralogous member obtain a new function after gene duplication), or subfunctionalization (where each paralog retains part of its original ancestral function) of duplicated genes by a series of synonymous and/or non-synonymous mutation during evolution[55]. In all, the majority of GrTCPs are evolutionarily conserved, while the variation of exon/intron distribution and motif composition in certain paralogous pairs of GrTCPs suggest that some members of TCP family in *G. raimondii* are functionally diversified through differential expansion. The diversification of GrTCPs may contribute to the unique functions of TCP transcription factors in cotton, for example controlling cotton fiber initiation and development. It has been reported that cotton transcription factor TCP14 was predominantly in cotton fiber cell, particularly at the stage of cotton fiber initiation and elongation[56]. Another study demonstrated that cotton TCP transcription factor regulated fiber

and root hair development by regulating jasmonic acid biosynthesis and response as well as ethylene signaling pathway[16]

References

1 Calkhoven, C. F. & Ab, G. Multiple steps in the regulation of transcription-factor level and activity. *Biochem. J.* **317 (Pt 2)**, 329-342 (1996).

2 Latchman, D. S. Transcription factors: an overview. *Int. J. Biochem. Cell Biol.* **29**, 1305-1312 (1997).

3 Schwechheimer, C., Zourelidou, M. & Bevan, M. W. Plant transcription factor studies. *Annu. Rev. Plant Phys.* **49**, 127-150 (1998).

4 Riechmann, J. L. *et al. Arabidopsis* transcription factors: Genome-wide comparative analysis among eukaryotes. *Science* **290**, 2105-2110 (2000).

5 Wang, K. *et al.* The draft genome of a diploid cotton Gossypium raimondii. *Nat. Genet.* **44**, 1098-1103, doi:10.1038/ng.2371 (2012).

6 Wray, G. A. *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**, 1377-1419, doi:10.1093/molbev/msg140 (2003).

7 Perez-Rodriguez, P. *et al.* PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res.* **38**, D822-827, doi:10.1093/nar/gkp805 (2010).

8 Cubas, P., Lauter, N., Doebley, J. & Coen, E. The TCP domain: a motif found in proteins regulating plant growth and development. *Plant J.* **18**, 215-222 (1999).

9 Martin-Trillo, M. & Cubas, P. TCP genes: a family snapshot ten years later. *Trends Plant Sci.* **15**, 31-39, doi:10.1016/j.tplants.2009.11.003 (2010).

10 Kosugi, S. & Ohashi, Y. PCF1 and PCF2 specifically bind to cis elements in the rice proliferating cell nuclear antigen gene. *Plant Cell* **9**, 1607-1619 (1997).

11 Yao, X., Ma, H., Wang, J. & Zhang, D. Genome-Wide Comparative Analysis and Expression Pattern of TCP Gene Families in *Arabidopsis thaliana* and *Oryza sativa*. *J. Integr. Plant Biol.* **49**, 885-897, doi:10.1111/j.1744-7909.2007.00509.x (2007).

12 Lupas, A., Vandyke, M. & Stock, J. Predicting Coiled Coils from Protein Sequences. *Science* **252**, 1162-1164 (1991).

13 Kosugi, S. & Ohashi, Y. DNA binding and dimerization specificity and potential targets for the TCP protein family. *Plant J.* **30**, 337-348 (2002).

14 Viola, I. L., Uberti Manassero, N. G., Ripoll, R. & Gonzalez, D. H. The *Arabidopsis* class I TCP transcription factor AtTCP11 is a developmental regulator with distinct DNA-binding properties due to the presence of a threonine residue at position 15 of the TCP domain. *Biochem. J.* **435**, 143-155, doi:10.1042/BJ20101019 (2011).

15 Schommer, C. *et al.* Control of jasmonate biosynthesis and senescence by miR319 targets. *PLoS Biol.* **6**, e230, doi:10.1371/journal.pbio.0060230 (2008).

16 Hao, J. *et al.* GbTCP, a cotton TCP transcription factor, confers fibre elongation and root hair development by a complex regulating system. *J. Exp. Bot.* **63**, 6267-6281, doi:10.1093/jxb/ers278 (2012).

17 Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680 (1994).

18 Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116-120, doi:10.1093/nar/gki442 (2005).

19 Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876-4882 (1997).
20 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725-2729, doi:10.1093/molbev/mst197 (2013).
21 Yang, S., Zhang, X., Yue, J. X., Tian, D. & Chen, J. Q. Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genomics.* **280**, 187-198, doi:10.1007/s00438-008-0355-0 (2008).
22 Gu, Z., Cavalcanti, A., Chen, F. C., Bouman, P. & Li, W. H. Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Mol. Biol. Evol.* **19**, 256-262 (2002).
23 Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452, doi:10.1093/bioinformatics/btp187 (2009).
24 Guo, A. Y., Zhu, Q. H., Chen, X. & Luo, J. C. [GSDS: a gene structure display server]. *Yi Chuan* **29**, 1023-1026 (2007).
25 Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369-373, doi:10.1093/nar/gkl198 (2006).
26 de Castro, E. *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* **34**, W362-365, doi:10.1093/nar/gkl124 (2006).
27 Aguilar-Martinez, J. A. & Sinha, N. Analysis of the role of *Arabidopsis* class I TCP genes AtTCP7, AtTCP8, AtTCP22, and AtTCP23 in leaf development. *Front. Plant Sci.* **4**, 406, doi:10.3389/fpls.2013.00406 (2013).
28 Takeda, T. *et al.* The OsTB1 gene negatively regulates lateral branching in rice. *Plant J.* **33**, 513-520 (2003).
29 Aguilar-Martinez, J. A., Poza-Carrion, C. & Cubas, P. *Arabidopsis* BRANCHED1 acts as an integrator of branching signals within axillary buds. *Plant Cell* **19**, 458-472, doi:10.1105/tpc.106.048934 (2007).
30 Palatnik, J. F. *et al.* Control of leaf morphogenesis by microRNAs. *Nature* **425**, 257-263, doi:10.1038/nature01958 (2003).
31 Kieffer, M., Master, V., Waites, R. & Davies, B. TCP14 and TCP15 affect internode length and leaf shape in *Arabidopsis*. *Plant J.* **68**, 147-158, doi:10.1111/j.1365-313X.2011.04674.x (2011).
32 Pagnussat, G. C. *et al.* Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. *Development* **132**, 603-614 (2005).
33 Sarvepalli, K. & Nath, U. Hyper-activation of the TCP4 transcription factor in *Arabidopsis* thaliana accelerates multiple aspects of plant maturation. *Plant J.* **67**, 595-607 (2011).
34 Takeda, T. *et al.* RNA interference of the *Arabidopsis* putative transcription factor TCP16 gene results in abortion of early pollen development. *Plant Mol. Biol.* **61**, 165-177 (2006).

35 Guo, Z. X. *et al.* TCP1 Modulates Brassinosteroid Biosynthesis by Regulating the Expression of the Key Biosynthetic Gene DWARF4 in *Arabidopsis* thaliana. *Plant Cell* **22**, 1161-1173 (2010).
36 Yanai, O., Shani, E., Russ, D. & Ori, N. Gibberellin partly mediates LANCEOLATE activity in tomato. *Plant J.* **68**, 571-582 (2011).
37 Hammani, K. *et al.* An *Arabidopsis* Dual-Localized Pentatricopeptide Repeat Protein Interacts with Nuclear Proteins Involved in Gene Expression Regulation. *Plant Cell* **23**, 730-740 (2011).
38 Giraud, E. *et al.* TCP Transcription Factors Link the Regulation of Genes Encoding Mitochondrial Proteins with the Circadian Clock in *Arabidopsis* thaliana. *Plant Cell* **22**, 3921-3934 (2010).
39 Pruneda-Paz, J. L., Breton, G., Para, A. & Kay, S. A. A Functional Genomics Approach Reveals CHE as a Component of the *Arabidopsis* Circadian Clock. *Science* **323**, 1481-1485 (2009).
40 Hubbard, L., McSteen, P., Doebley, J. & Hake, S. Expression patterns and mutant phenotype of teosinte branched1 correlate with growth suppression in maize and teosinte. *Genetics* **162**, 1927-1935 (2002).
41 Uberti-Manassero, N. G., Lucero, L. E., Viola, I. L., Vegetti, A. C. & Gonzalez, D. H. The class I protein AtTCP15 modulates plant development through a pathway that overlaps with the one affected by CIN-like TCP proteins. *J. Exp. Bot.* **63**, 809-823, doi:10.1093/jxb/err305 (2012).
42 Koyama, T., Ohme-Takagi, M. & Sato, F. Generation of serrated and wavy petals by inhibition of the activity of TCP transcription factors in *Arabidopsis* thaliana. *Plant Signal. Behav.* **6**, 697-699 (2011).
43 Tatematsu, K., Nakabayashi, K., Kamiya, Y. & Nambara, E. Transcription factor AtTCP14 regulates embryonic growth potential during seed germination in *Arabidopsis* thaliana. *Plant J.* **53**, 42-52, doi:10.1111/j.1365-313X.2007.03308.x (2008).
44 Rueda-Romero, P., Barrero-Sicilia, C., Gomez-Cadenas, A., Carbonero, P. & Onate-Sanchez, L. *Arabidopsis* thaliana DOF6 negatively affects germination in non-after-ripened seeds and interacts with TCP14. *J. Exp. Bot.* **63**, 1937-1949, doi:10.1093/jxb/err388 (2012).
45 Doebley, J., Stec, A. & Hubbard, L. The evolution of apical dominance in maize. *Nature* **386**, 485-488 (1997).
46 Luo, D., Carpenter, R., Vincent, C., Copsey, L. & Coen, E. Origin of floral asymmetry in Antirrhinum. *Nature* **383**, 794-799 (1996).
47 Crawford, B. C. W., Nath, U., Carpenter, R. & Coen, E. S. CINCINNATA controls both cell differentiation and growth in petal lobes and leaves of antirrhinum. *Plant Physiol.* **135**, 244-253 (2004).
48 Nath, U., Crawford, B. C. W., Carpenter, R. & Coen, E. Genetic control of surface curvature. *Science* **299**, 1404-1407 (2003).
49 Lewis, J. M. *et al.* Overexpression of the maize Teosinte Branched1 gene in wheat suppresses tiller development. *Plant Cell Rep.* **27**, 1217-1225 (2008).
50 Li, C. X., Potuschak, T., Colon-Carmona, A., Gutierrez, R. A. & Doerner, P. *Arabidopsis* TCP20 links regulation of growth and cell division control pathways. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12978-12983 (2005).

51 *Arabidopsis* Genome, I. Analysis of the genome sequence of the flowering plant *Arabidopsis* thaliana. *Nature* **408**, 796-815, doi:10.1038/35048692 (2000).
52 Zhang, J. Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**, 292-298, doi:10.1016/s0169-5347(03)00033-8 (2003).
53 Flagel, L. E. & Wendel, J. F. Gene duplication and evolutionary novelty in plants. *New Phytol.* **183**, 557-564, doi:10.1111/j.1469-8137.2009.02923.x (2009).
54 Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant cell* **16**, 1667-1678, doi:10.1105/tpc.021345 (2004).
55 Prince, V. E. & Pickett, F. B. Splitting pairs: the diverging fates of duplicated genes. *Nat. Rev. Genet.* **3**, 827-837, doi:10.1038/nrg928 (2002).
56 Wang, M. Y. *et al.* The cotton transcription factor TCP14 functions in auxin-mediated epidermal cell differentiation and elongation. *Plant Physiol.* **162**, 1669-1680, doi:10.1104/pp.113.215673 (2013).

Table 1.1: TCP gene family in *Gossypium raimondii*

Gene Name	Gene symbol	Length(aa)	MW(Da)	pI	Chr. Location
GrTCP1	Gorai.001G200400.1	398	43633.3	9.3687	Chr01: 36491376 - 36493627
GrTCP2	Gorai.009G153900.1	410	45001.5	8.2869	Chr09: 11754631 - 11761100
GrTCP3	Gorai.002G064500.1	444	48127.5	6.9763	Chr02: 7559171 - 7562527
GrTCP4	Gorai.009G373000.1	401	43857.1	6.7323	Chr09: 50501688 - 50504320
GrTCP5	Gorai.008G199700.1	327	36256.9	6.1773	Chr08: 48462197 - 48464598
GrTCP6	Gorai.011G086900.1	300	31920.4	8.5291	Chr11: 9062285 - 9064732
GrTCP7a	Gorai.005G211900.1	257	26592.8	10.1465	Chr05: 59315620 - 59319891
GrTCP7b	Gorai.013G068600.1	256	26423.6	10.3339	Chr13: 7948087 - 7951341
GrTCP7c	Gorai.008G147800.1	243	25363.6	10.6333	Chr08: 40133983 - 40134714
GrTCP8	Gorai.012G166500.1	488	51182	8.0412	Chr12: 33695542 - 33697571
GrTCP9a	Gorai.007G094200.1	338	35440.4	9.539	Chr07: 6915362 - 6917586
GrTCP9b	Gorai.004G206900.1	345	36349.7	8.3417	Chr04: 53747390 - 53748427
GrTCP10	Gorai.013G172800.1	409	44213.7	7.3274	Chr13: 45973805 - 45976133
GrTCP11	Gorai.006G165300.1	270	29631.8	9.285	Chr06: 42514390 - 42515568
GrTCP12	Gorai.008G186800.1	501	55904.8	7.1546	Chr08: 46771143 - 46773319
GrTCP13a	Gorai.012G048500.1	309	34162.3	8.9496	Chr12: 6242356 - 6244652
GrTCP13b	Gorai.006G009800.1	285	32029.8	7.9301	Chr06: 2180268 - 2182165
GrTCP14a	Gorai.007G036800.1	395	42222.1	7.4496	Chr07: 2525369 - 2527057
GrTCP14b	Gorai.001G072200.1	409	43090.4	8.531	Chr01: 7309330 - 7311843
GrTCP14c	Gorai.008G192400.1	401	43363.7	7.2316	Chr08: 47573318 - 47575119
GrTCP15a	Gorai.008G181600.1	344	37542.6	9.1122	Chr08: 45925248 - 45927562
GrTCP15b	Gorai.N023400.1	365	39735.1	9.737	scaffold_505: 23 - 1120
GrTCP15c	Gorai.013G084500.1	352	38234.7	9.8567	Chr13: 12166527 - 12169530
GrTCP16	Gorai.N023500.1	216	24069.9	9.9096	scaffold_505: 5633 - 6280
GrTCP17	Gorai.001G076700.1	266	30361.1	8.054	Chr01: 7898275 - 7903253
GrTCP18a	Gorai.007G007500.1	324	37215.3	8.8024	Chr07: 600999 - 602878
GrTCP18b	Gorai.008G285300.1	361	40855.6	8.2647	Chr08: 56105392 - 56106838
GrTCP19a	Gorai.006G197000.1	337	36419.8	7.118	Chr06: 45427588 - 45429051
GrTCP19b	Gorai.008G243000.1	388	41371.5	8.7407	Chr08: 52845411 - 52846971
GrTCP20a	Gorai.012G084600.1	300	31763.5	9.1033	Chr12: 14283464 - 14287212
GrTCP20b	Gorai.006G043800.1	298	31463.1	9.9629	Chr06: 13026618 - 13028750
GrTCP20c	Gorai.001G273300.1	334	35872.1	10.08	Chr01: 55023721 - 55025665
GrTCP20d	Gorai.008G157300.1	298	31748.5	10.3948	Chr08: 42005542 - 42007175
GrTCP21	Gorai.013G268200.1	196	21060.7	8.746	Chr13: 58033580 - 58034170
GrTCP22	Gorai.002G215000.1	549	57867.3	7.2398	Chr02: 56433132 - 56435832
GrTCP23	Gorai.009G289000.1	421	44684.3	7.4635	Chr09: 24825613 - 24826878
GrTCP24	Gorai.011G046000.1	463	50079.8	7.5515	Chr11: 3523288 - 3528930
GrTCP25	Gorai.009G398700.1	435	48308	6.4093	Chr09: 56724136 - 56726696

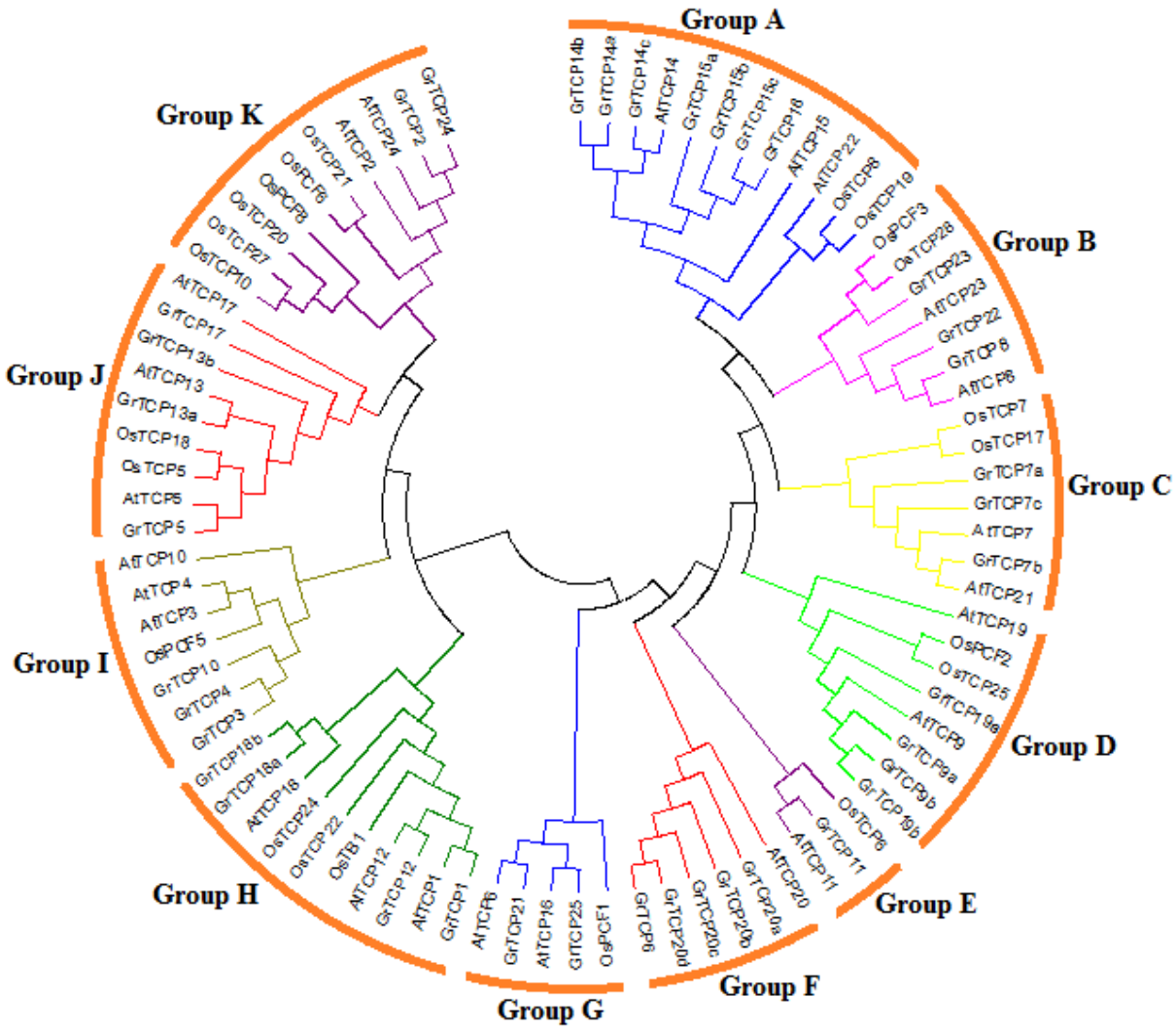


Figure 1.1. Phylogenetic relationships of TCP transcription factors from *Gossypium ramondii*, *Arabidopsis* and rice. The unrooted phylogenetic tree was constructed using MEGA 6.0 by Neighbor-Joining method and the bootstrap test was performed with 1,000 iterations. The eleven subclades are indicated with different colors.

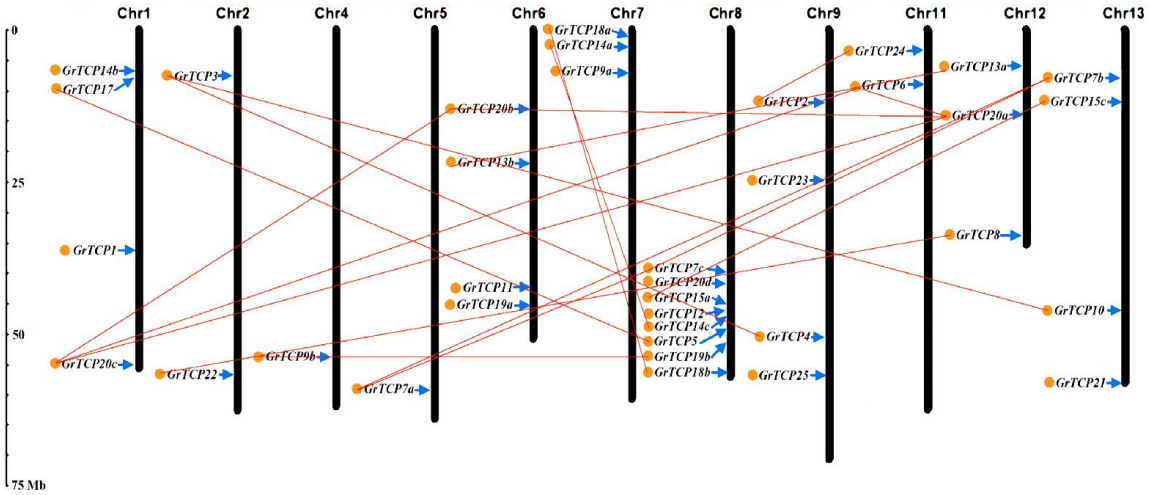


Figure 1.2. Chromosomal distribution and gene duplication of TCP genes in *G. raimondii*. The scale is in megabases (Mb). The chromosome numbers are indicated at the top of each chromosome. The paralogous TCP genes are connected with a red line.

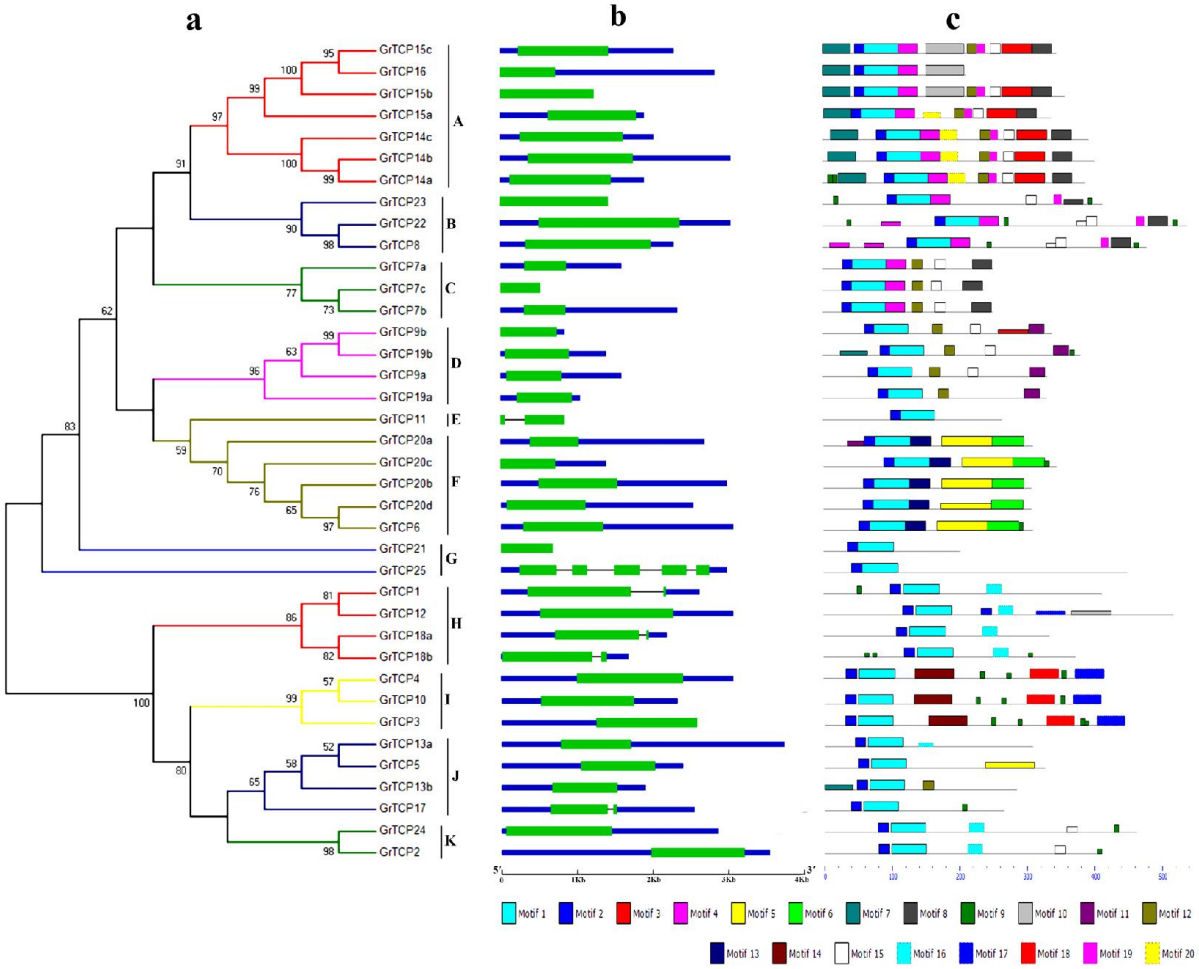


Figure 1.3. Phylogenetic analysis, gene structure and conserved motifs of TCP family in *Gossipium raimondii*. a. The phylogenetic tree of all TCP transcription factors in *G. raimondii* was constructed using Neighbor-Joining method and the bootstrap test was performed with 1,000 iterations. Bootstrap values higher than50% support are displayed. **b.** The exon/intron organization of TCP genes of *G. raimondii*. The blue lines represent 5'-UTR or 3'-UTR, green boxes represent exons and black lines indicate introns. **c.** The conserved protein motifs in the TCP family were identified using MEME program. Each motif is indicated with a specific color.

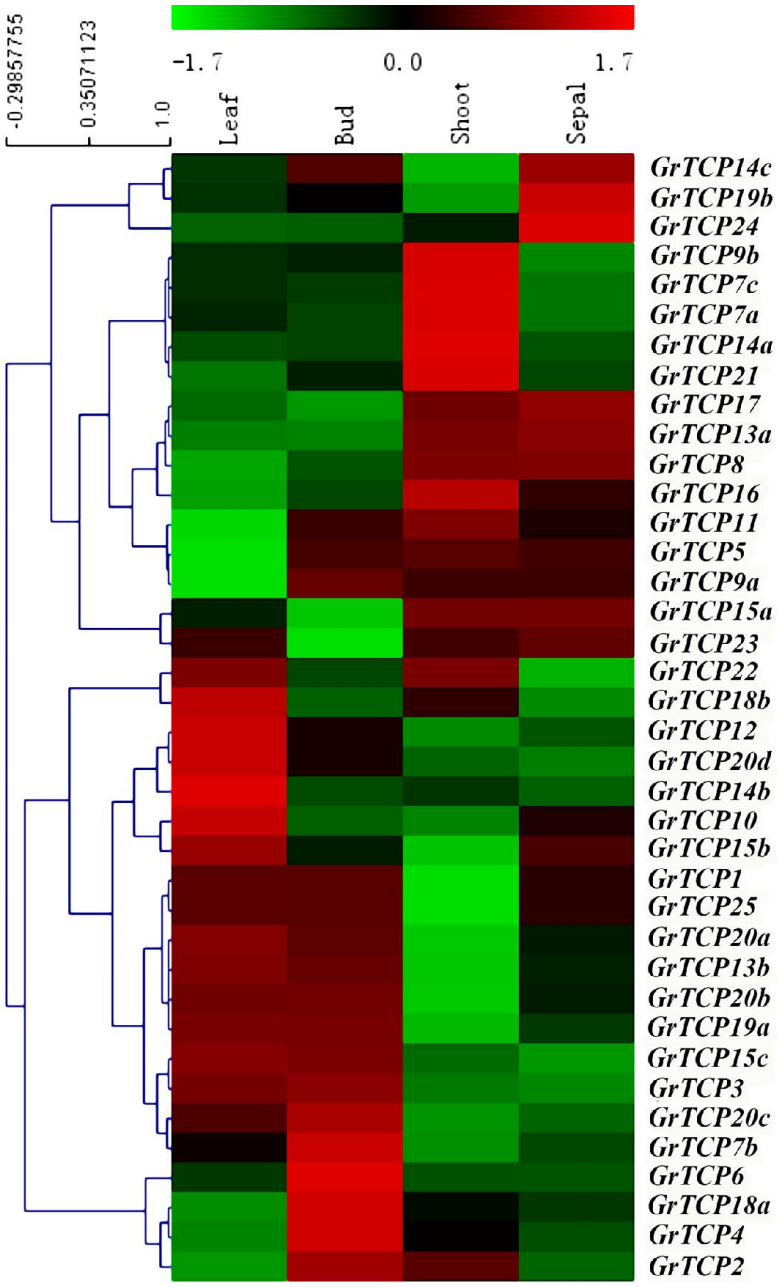


Figure 1.4. Heatmap representation for expression patterns of *G. raimondii* TCP genes across different tissues. The expression profile data of GrTCP genes in leaf, bud, shoot and sepal were obtain through quantitative real-time PCR.

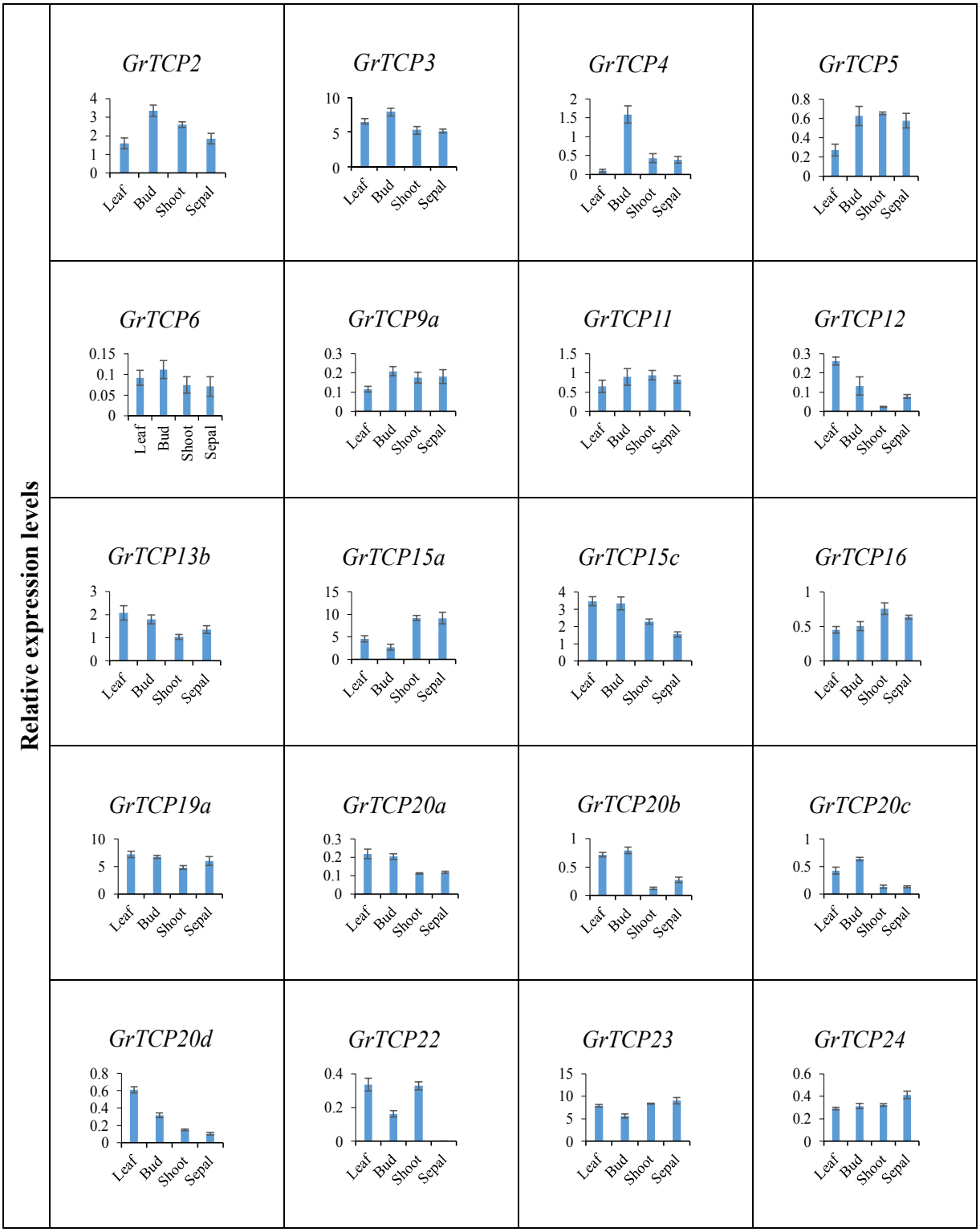


Figure 1.5. Expression profiles of 20 GrTCP genes across different tissues. The y-axis represent the relative expression levels of GrTCP genes against reference gene *TuA11*. Error bars indicate standard deviation for three replicates.

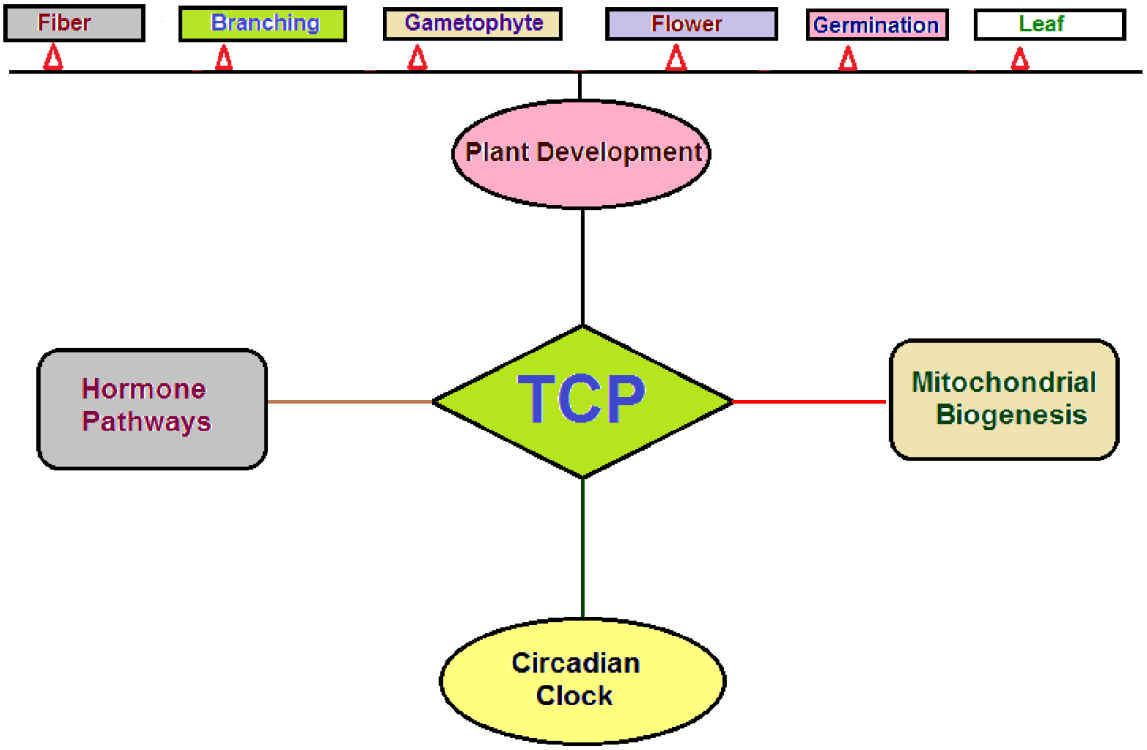
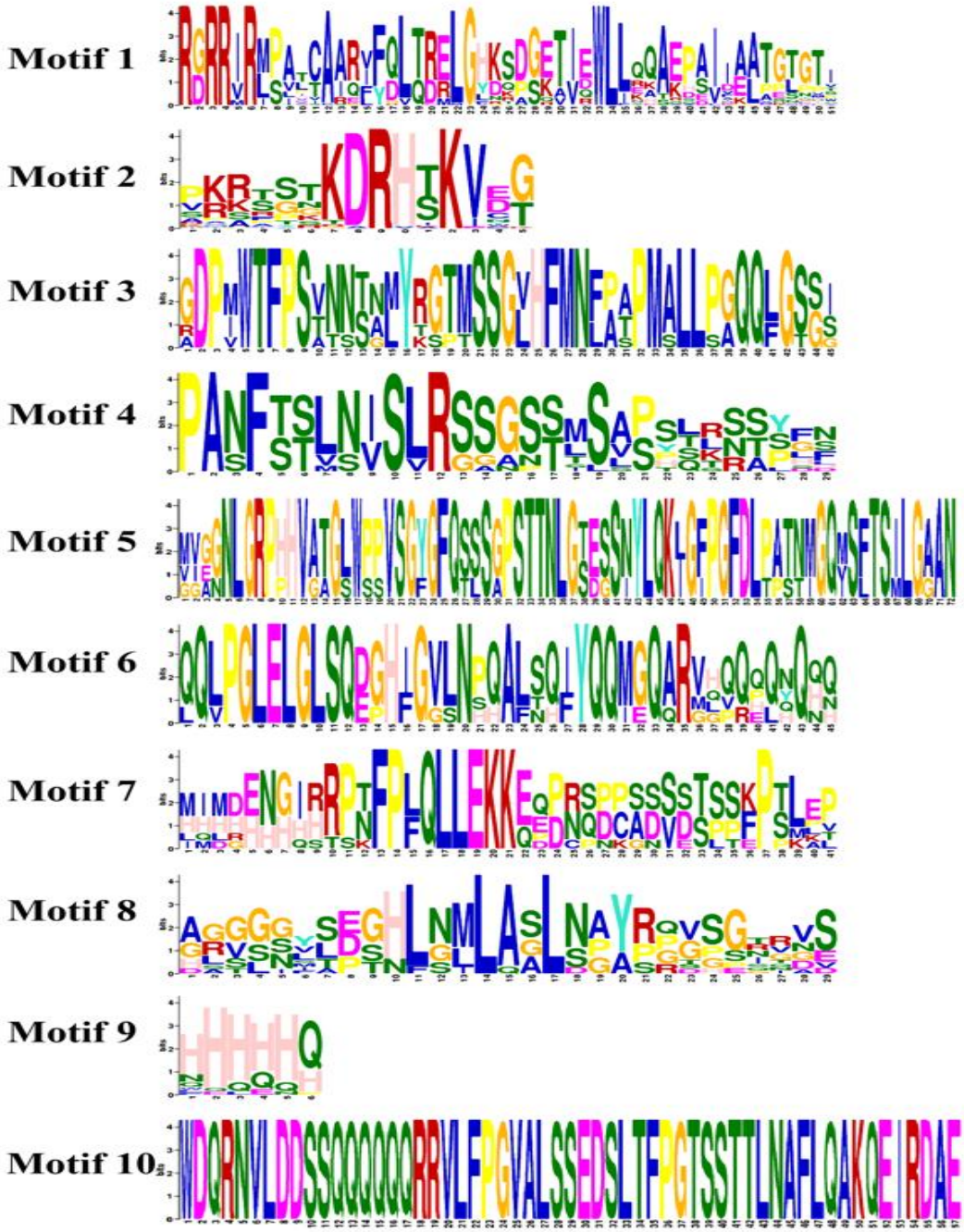
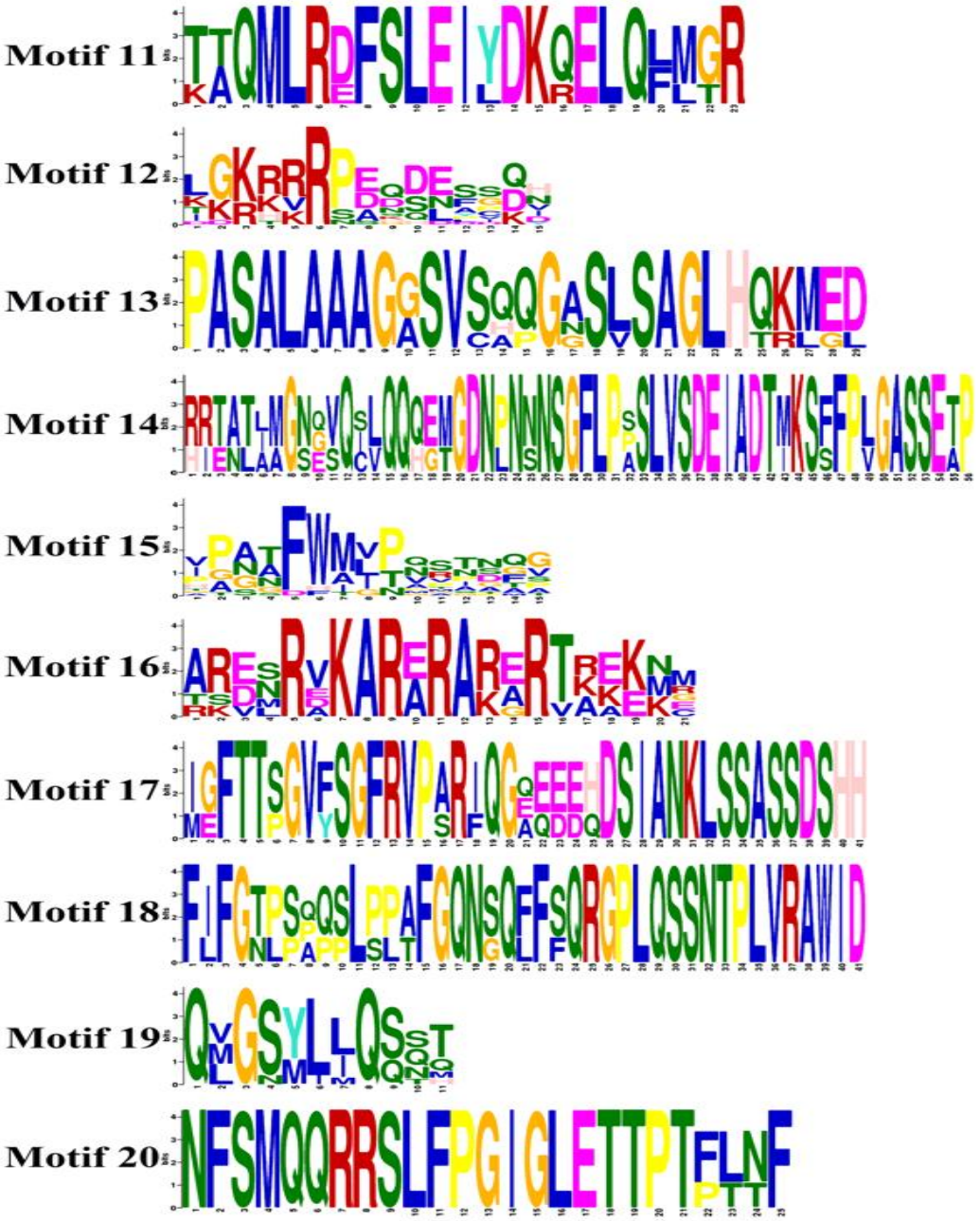


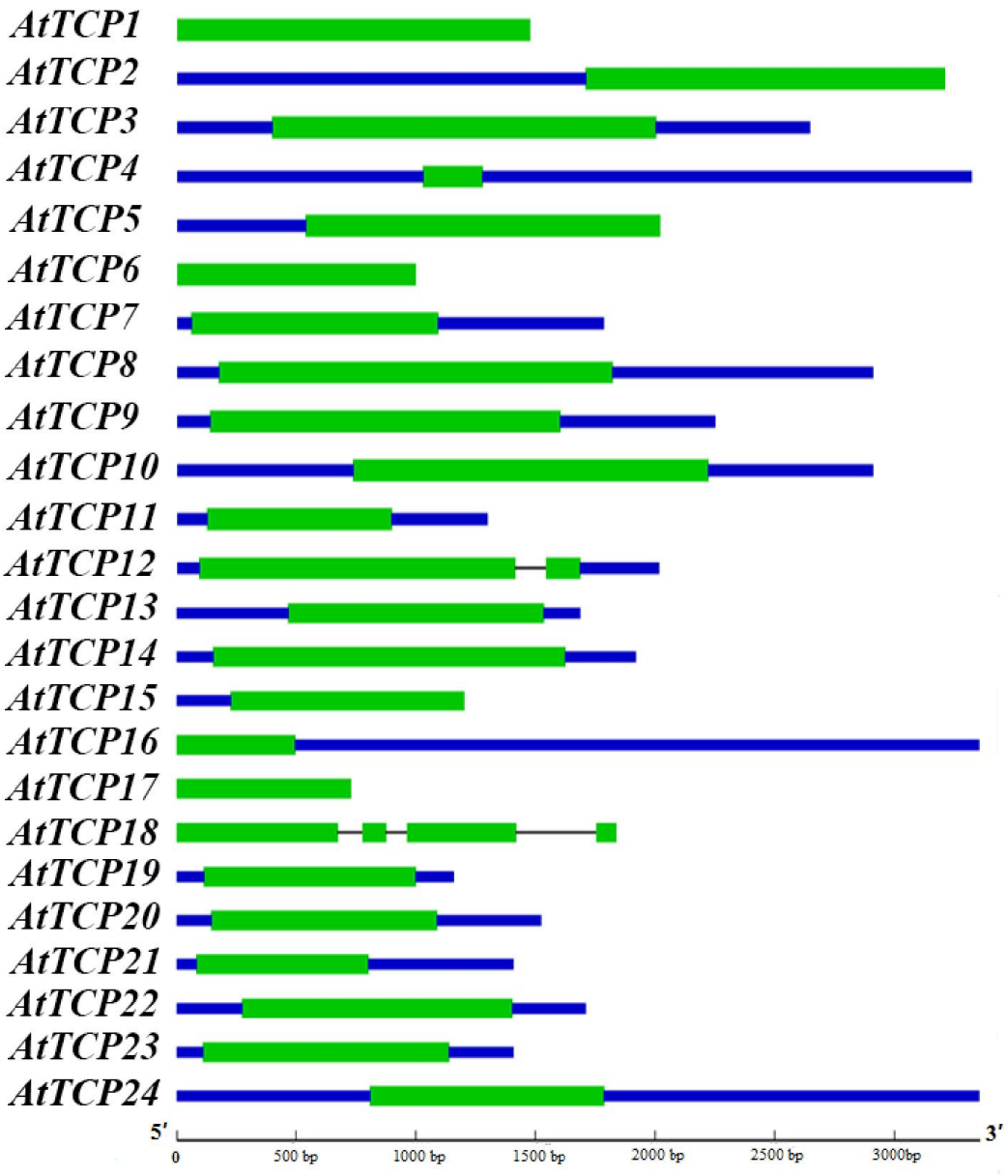
Figure 1.6. TCP transcription factors play important role in the multiple biological process during plant growth and development.


Motif 1
Motif 2
Motif 3
Motif 4
Motif 5
Motif 6
Motif 7
Motif 8
Motif 9
Motif 10

(continued)



Supplementary Figure 1.1. **Sequence logos for conserved motifs identified in GrTCP proteins**. The total height of the letters stands for the information content of the position in bits and numbers on the horizontal axis represent the sequence positions in the motifs.



Supplementary Figure 1.2. Exon/intron structure of *Arabidopsis* TCP genes. The blue lines represent 5'-UTR or 3'-UTR, green boxes represent exons and black lines indicate introns.

CHAPTER 3:Comprehensive Analysis of TCP Transcription Factor Family in Cotton (*Gossypium arboreum* L.)

Abstract

TCP proteins are plant-specific transcription factors implicated to perform a variety of physiological functions during plant growth and development. In the current study, we performed for the first time the comprehensive analysis of TCP gene family in a diploid cotton species, *Gossypium arboreum*, including phylogenetic analysis, chromosome location, gene duplication status, gene structure and conserved motif analysis, as well as expression profiles in fiber at different developmental stages. Our results showed that *G. arboreum* contains 36 TCP genes, distributing across all of the thirteen chromosomes. GaTCPs within the same subclade of the phylogenetic tree shared similar exon/intron organization and motif composition. In addition, both segmental duplication and whole-genome duplication contributed significantly to the expansion of GaTCPs. Many these TCP transcription factor genes are specifically expressed in cotton fiber during different developmental stages, including cotton fiber initiation and early development. This suggests that TCP genes may play important roles in cotton fiber development.

Keywords: Cotton, *Gossypium arboreum*, TCP, fiber

Introduction

TCP proteins constitute a family of plant-specific transcription factors widely distributed in angiosperms [1,2]. The TCP gene family was termed after its founding members: TEOSINTE BRANCHED 1 in *Zea mays*, CYCLOIDEA in *Antirrhinum majus* and PCF in *Oryza sativa* [1]. Since its initial identification and characterization in 1999, the TCP family has become one of the focuses of plant studies due to its importance in the evolution and developmental control of plant form [2]. TCP proteins are defined by a 59-residue-long basic helix-loop-helix (bHLH) structure called TCP domain, which provides this family with the ability to bind GC-rich DNA sequence motifs [2]. According to the secondary structure prediction, the basic region of TCP domain is followed by two helices separated by a loop [2]. In addition, phylogenetic analysis showed that TCP proteins can be classified into two subfamilies based on their DNA binding domain structure [2]. To date, more than 20 TCP family members have been identified in a number of monocot and eudicot plants, such as *Arabidopsis* [3], *Oryza sativa* [4], *Vitis vinifera* and *Populus trichocarpa* [5].

In plants, the TCP transcription factor family has been implicated to perform a variety of physiological functions during plant growth and development, such as branching, regulation of the circadian clock, seed germination, gametophyte development, hormone pathways, leaf development, mitochondrial biogenesis, flower development and cell cycle regulation [2,6-12]. Recently, evidences indicated that TCP proteins also play a significant role in fiber development [13,14], which makes it necessary to identify and characterize TCP family members in cotton, one of the most important economic crops and natural fiber sources all over the world [15].

Cotton comprises both diploid and tetraploid species, belonging to the *Gossypium* genus. The most commonly cultivated cotton species for fiber and oil production is upland cotton (*Gossypium hirsutum*), an AD tetraploid evolved from A-genome diploids such as *G. arboreum* and D-genome diploids like *G. raimondii* at around 1-2 million years ago [16]. Up to now, only two TCP family members have been functionally characterized in cotton, suggesting that TCP genes may play key roles in fiber development [13,14]. Therefore, there is an urgent need to perform a

genome wide analysis of this family in cotton. The recent completion of the sequencing of *G. arboreum* genome allowed us to characterize all cotton TCP genes.

In the current study, we performed for the first time the comprehensive analysis of TCP gene family in *G. arboreum*, including phylogenetic analysis, chromosome location, gene duplication status, gene structure and conserved motif analysis, as well as tissue specific expression profiles. Our findings will lay solid foundation to better understand the function and evolutionary history of GaTCPs, and will help further investigation of the detailed molecular and biological functions of TCP members in cotton.

Materials and methods

Identification of TCP genes and proteins

The genome sequence of *G. arboreum* was downloaded from the Cotton Genome Project (CGP) (http://cgp.genomics.org.cn/). To identify the TCP family in *G. arboretum*, the profile hidden Markov Models of TCP domain (PF03634) downloaded from Pfam was used as query to run Hmmsearch against the *G. arboretum* genome (P-value = 0.0001). The candidate TCP genes were further aligned to remove redundant sequences [17]. Subsequently, the TCP sequences were manually inspected with ScanProsite to confirm the presence of the conserved TCP domain [18]. The TCP gene and protein sequences from *Theobroma cacao*, *Vitis vinifera*, *Arabidopsis thaliana* and *Oryza sativa* were retrieved from PlantTFDB plant transcription factor database, while the GrTCP sequences were obtained from previous studies [19].

Phylogenetic analysis

Cluster X program was employed to perform multiple sequence alignments with default parameters [20]. Unrooted phylogenetic trees were subsequently constructed using the Neighbor-Joining (NJ) method implemented in the MEGA 6.0 software with JTT model and pairwise gap deletion option [21]. The bootstrap analysis was conducted with 1000 iterations.

Chromosomal location and gene duplication

The physical location data of GaTCP genes were retrieved from *G. arboreum* genome. Mapping of these GaTCP genes was then performed using MapInspect software. Gene duplication was defined according to the criteria described in previous studies: the aligned region of two sequences covers over 70% of the longer sequence and the similarity of the aligned region is over 70% [22,23]. In addition, the DnaSp software [24] was employed to calculate Ka (nonsynonymous substitution rate) and Ks (synonymous substitution rate), which was further used to estimate the date of duplication events with the formula $T = Ks/2\lambda$, assuming clock-like rate (λ) of 1.5 synonymous substitutions per 10^8 years for cotton [25].

Gene structure and conserved motif

The exon/intron organizations of GaTCPs were inferred through comparison of genomic sequences and CDS sequences in the gene structure display server [26]. The program MEME [27] was employed to identify conserved motifs in GaTCPs with the following parameters: the optimum width of motif, 6-250; the maximum number of motif, 20; the number of repetitions, any. In addition, motif annotation was performed using the program InterProScan [28].

RNA isolation and Real-time quantitative RT-PCR analysis

Total RNA was isolated from *G. arboreum* leaves, sepals and fibers at -2, 0, 2, 5 and 10 days post anthesis (DPA) using the mirVanaTM miRNA Isolation Kit (Ambion, USA). Subsequently, the NanoDrop ND-1000 Spectrophotometer was employed to determine RNA concentration and quality. cDNA was then synthesized from 1 μg of total RNA with poly-T primers using the TaqMan® MicroRNA Reverse Transcription Kit (Applied Biosystems, USA). RT-qPCR was later conducted on 7300 Real-Time PCR System (Applied Biosystems, USA) according to the manufacture's protocol. The amplification parameters were as follows: enzyme activation at 95°C for 10 min, 45 cycles of denaturation at 95°C for 15s and annealing/elongation at 60 °C for 60 s. The relative expression levels was calculated according to previous studies [19]. A reference genes *SAD1* was used to normalize the expression values. There were three biological replicates

and each biological replicate was run three times. Finally, the software MultiExperiment Viewer was used to construct heatmap representation for expression patterns.

Results

Identification of the TCP gene family in G. arboretum

The TCP transcription factor family is featured by a highly conserved TCP domain at the N-terminus. To identify this family in *G. arboretum*, the profile hidden Markov Models of TCP domain (PF03634) downloaded from Pfam was used as query to run Hmmsearch against the *G. arboretum* genome. As a result, 58 putative TCP genes were identified initially. After the removal of 22 redundant sequences based on multiple sequence alignment, 36 candidate TCP genes sequences were manually inspected with ScanProsite to confirm the existence of the conserved TCP domain. As expected, they all contained the TCP domain, suggesting that they were members of the TCP gene family. The 36 TCP genes were further named as *GaTCP1* to *GaTCP25* according to *Arabidopsis* TCP nomenclature suggestions. Table 2.1 summarized their gene symbols, corresponding gene names, sequence length, molecular weights, isoelectric points and chromosome location.

Evolutionary analysis of the TCP transcription factor family

In order to explore the evolutionary relationships of the TCP transcription factor family, an unrooted phylogenetic tree was generated using the full length TCP protein sequences from *G. arboreum*, *G. raimondii*, *Theobroma cacao*, *Vitis vinifera*, *Arabidopsis thaliana* and *Oryza sativa*. As illustrated in the Neighbor-Joining phylogenetic tree (Figure 2.1), the TCP transcription factor family was classified into ten distinct subgroups designated as Group A to Group J. Group E is composed of 33 members, constituting the largest clade among all subgroups, while Group G is the second largest clade, consisting of 21 TCP protein sequences. The smallest clade is Group D, which contains only four TCP proteins. Generally speaking, all subgroups contain at

least five plant species except Group D, exhibiting an interspersed distribution. This may imply that the divergence of these species took place after the TCP transcription factor family expanded. Noticeable, the TCPs from *G. arboretum* and *G. raimondii* formed quite a few clusters of homologs in all subfamilies due to their high sequence similarity and close evolutionary relationship that the divergence of the two cotton species occurred only 2–13 million years ago [29]. In addition, the cotton TCPs (GrTCPs and GaTCPs) were more closely allied to TCP proteins from *T. cacao* than from other plant species, consistent with the fact that *G. arboreum*, *G. raimondii* and *T. cacao* originated from a common ancestor 18–58 million years ago [29]. Moreover, Group B and Group D did not contain *O. sativa* TCP proteins, suggesting that the TCP family members in the two subgroups were either lost in *O. sativa* or obtained after the divergence of monocots and eudicots. Additionally, the phylogenetic analysis also showed that the TCP members from different plant species were not evenly distributed in some subgroups. GaTCPs, for example, were overrepresented than AtTCPs in Group C and Group E, in which GaTCPs were over two times the number of AtTCPs. This may indicate that the TCPs went through differential expansion in *G. arboreum* and in *Arabidopsis*.

Chromosomal distribution and gene duplication

The 36 *G. arboretum* genes were mapped onto chromosomes in order to elucidate their chromosomal distribution and gene duplication status. As shown in Figure 2.2, the 36 GaTCPs were scattered throughout all 13 chromosomes of *G. arboretum*. Chromosome 6 had the highest number of five TCP genes, followed by Chromosome 10 and Chromosome 13 with four TCP genes. The lowest number of GaTCPs were observed in Chromosome 4 with one TCP gene. In general, the TCP genes were more evenly distribute across *G. arboretum* chromosomes than that in *G. raimondii* [19].

In addition, the gene duplication events were further investigated to reveal the expansion mechanism of the TCP gene family in *G. arboretum*. The criteria described in previous studies were employed to identify paralogous genes pairs. As a result, 13 pairs of putative paralogous TCP genes were found in *G. arboretum* with high gene and protein sequence identity and similarity, accounting for about 72% of the whole GaTCP gene family. As illustrated in Figure

2.2, all of the gene pairs were distributed on different chromosomes, while no tandem duplication events could be observed, suggesting that segmental duplications contributed a lot to the amplification of the GaTCP gene family. Additionally, in order to gain more insights of the evolutionary history of the GaTCP gene family, the DnaSP program was used to calculate the approximate dates of duplication events, dating the segmental duplication of GaTCPs between 11.56 Mya (million years ago) to 37.12 Mya, with an average of 19.27 Mya.

Gene structure and conserved motifs

To get a better understanding of the diversification of the TCP genes in *G. arboretum*, the exon/intron organization and conserved motifs of GaTCPs were analyzed. A new Neighbor-Joining phylogenetic tree was constructed using the protein sequences of GaTCPs, dividing the TCP family into eight subclades. As shown in Figure 2.3, over 80% of GaTCP genes were intronless, which was quite similar to the structure of *G. raimondii* TCP genes [19]. Generally speaking, most GaTCPs within the same subclades exhibited similar gene structure in terms of numbers and lengths of introns and exons. Subclade A and H, for instance, contained intronless genes with similar exon lengths. In contrast, great structure variants were observed in Subclade C. In addition, the MEME programs was used to predict motif composition, identifying twenty conserved motifs in GaTCPs. Subsequently, the program InterProScan was employed to annotate these motifs. The results showed that the only motif that hit for the database was the conserved TCP domain (the red motif), which was found in all GaTCPs. Moreover, similar to the exon/intron organization, members belonging to the same subclades also showed similar motif composition, indicating their functional similarities. Additionally, some motifs were only presented at specific subclades, suggesting that they may perform subclade specific functions.

Expression profiles of cotton TCP genes at different developmental stages

In order to shed light on the potential physiological functions of the TCP gene family in *G. arboretum* at different developmental stages as well as in cotton fiber development, their expression profiles were investigated using Real-Time Quantitative Reverse Transcription PCR on several different organs, including leaves, sepals and fibers at -2, 0, 2, 5 and 10 DPA. As

shown in Figure 2.4 and Figure 2.5, the majority of the TCP genes exhibited diverse expression profiles, while a few of them showed similar expression patterns. For example, a number of genes, including *GaTCP5*, *GaTCP8*, *GaTCP12*, *GaTCP13a*, *GaTCP14b*, *GaTCP15a*, *GaTCP15b*, *GaTCP19a*, *GaTCP21*, were exclusively highly expressed in fiber, while *GaTCP13b* and *GaTCP20b* were preferentially expressed in leaves or sepals at the high levels. Additionally, several TCP genes had high expression levels in all the tissues examined, such as *GaTCP6*, *GaTCP18b* and *GaTCP23*. Among those genes that were highly expressed in fibers, *GaTCP7a*, *GaTCP12*, *GaTCP13a* and *GaTCP24* were highly expressed at the fiber initiation stage (from -2 to 2 DPA), whereas the expression levels of *GaTCP5*, *GaTCP10*, *GaTCP14c* and *GaTCP15b* were significantly high at the fiber elongation stage (from 5 to 10 DPA). Remarkably, *GaTCP8* had extremely high expression levels at all the fiber developmental stages tested.

Discussion

The TCP transcription factor family plays important roles in many biological processes during plant growth and development, such as fiber development [13,14], seed germination [30,31], leaf development [32], hormone signal transduction [33] and flower development [32-34]. The recent availability of *G. arboreum* genome sequences [29] allowed us to perform a comprehensive analysis of this family in cotton, including phylogenetic analysis, chromosome location, gene duplication status, gene structure and conserved motif analysis, as well as tissue specific expression profiles.

Evolutionary conservation and divergence of the TCP gene family in cotton

The *G. arboreum* genome contains almost the same number of TCP genes as in *G. raimondii* [19] and have more than 50% more than in *Arabidopsis* [2], which is in consistency with the number of protein coding genes in each species [25,29,35]. It has been reported that *Arabidopsis* and *G. arboreum* evolved from a common ancestor at around 93 Mya and subsequently underwent paleopolyploidy events [25,36]. Our phylogenetic analysis showed that TCP genes in cotton and *Arabidopsis* might go through differential expansion caused by gene duplication, an important source of raw genetic materials to the evolution of complex plant systems [37,38]. This is supported

by the fact that the number of paralogous TCP gene pairs accounted for over 70% of the entire TCP gene family in *G. arboreum* and that segmental duplication is a predominant duplication event for GaTCPs. Previous studies indicated that gene duplication contributed to the amplification of gene family members on various scales, such as tandem duplication, segmental duplication and whole-genome duplication, and that the expansion of regulatory genes can hardly ever be achieved simply through single gene duplication alone [37-39], implying that genome duplication may also contribute to the amplification of the GaTCP gene family. According to a recent study, a recent and an ancient whole genome duplication event have occurred in *G. arboreum* at approximately 13–20 and 115–146 Mya, respectively [29]. Our results indicated that the average duplication date of GaTCPs was around 19.27 Mya, which is consistent with the recent whole genome duplication event. This suggests that genome duplication may also play a significant role in the expansion of the GaTCP gene family.

In addition to gene duplication status, differences in exon/intron organizations can also shed light on the evolutionary history of gene families. In this study, we compared the gene structure of GaTCPs with their homologous counterparts in *Arabidopsis* [19]. The results showed that eight of ten pairs of TCP genes shared conserved exon/intron distribution in terms of exon length and intron numbers, whereas two pairs displayed some extent of divergence. *AtTCP12*, for instance, contained one more intron than its counterpart *GaTCP12*. Such intron loss of gain may result from insertion/deletion events in the process of evolution [40].

According to previous reports, upland cotton (*G. hirsutum*), an AD tetraploid, evolved from A-genome diploids *G. arboreum* and D-genome diploids *G. raimondii* at around 1-2 Mya [16]. Due to high similarities between the A and D genomes in terms of gene sequence and genome organization, the TCP family members in *G. arboreum* and *G. raimondii* formed a lot of clusters of homologs in the phylogenetic tree and shared almost identical exon/intron structures as well as motif compositions [19]. Our phylogenetic analysis also showed that GaTCPs and GrTCPs were more closely allied to TCP proteins from *T. cacao*, a close relative of cotton in the *Malvaceae* family, than from other plant species, consistent with the fact that *G. arboreum*, *G. raimondii* and *T. cacao* originated from a common ancestor 33 Mya [29]. Since GaTCPs and GrTCPs duplicated at around 19.27 million years ago, these duplications happened after their divergence from *T.*

cacao and *Arabidopsis* but before the reunion of the A and D genome diploids that gave rise to upload cotton.

Functional divergence of the TCP gene family in cotton

It has been widely recognized that duplicated genes undergo one of the following evolutionary fates: pseudogenization (in which one copy becomes unexpressed or functionless), conservation of gene function (in which both copies maintain the same function), subfunctionalization (in which the ancestral function is subdivided between copies), and neofunctionalization (in which one copy acquires a new function) [37]. In the present study, the expression profiles of GaTCPs at different developmental stages were investigated to reveal their functional divergence during plant growth and development.

Our results showed that the majority of the paralogous GaTCP gene pairs exhibited differential expression profiles. *GaTCP7a*, for example, was preferentially expressed in fiber at the initiation stage, while its paralogous counterpart *GaTCP23* was highly expressed in fiber at both the initiation stage and the elongation stage. *GaTCP14c* had extremely high expression level in leaf and fiber at the elongation stage, whereas *GaTCP14b* was relatively highly expressed in fiber at the initiation stage. In general, the expression patterns of GaTCP genes imply that the TCP family may perform multiple physiological functions in *G. arboreum,* especially in fiber initiation and elongation. Remarkably, some of these findings have already been experimentally confirmed through analysis of mutant cotton species with reduced and/or overexpressed TCP activities. For instance, the expression level of *GaTCP15b* was significantly high in fiber at the elongation stage. Hao *et al.* demonstrated that *GbTCP*, the homologous counterpart of *GaTCP15b* in *G. barbadense*, confers cotton fiber elongation by regulating JA biosynthesis and response and other pathways using RNAi silencing technique [14]. In addition, Wang *et al.* demonstrated that the *GhTCP14* from upland cotton functions as a crucial regulator in auxin-mediated elongation of cotton fiber cells [13], which is in agreement with our result that *GaTCP14c* was highly expressed in fiber at the elongation stage. However, further functional analysis of GaTCPs are still needed in order to determine which evolutionary fates the duplicated GaTCP genes undergo during the process of sequence and functional evolution.

References

1 Cubas, P., Lauter, N., Doebley, J. & Coen, E. The TCP domain: a motif found in proteins regulating plant growth and development. *The Plant Journal* **18**, 215-222, doi:10.1046/j.1365-313X.1999.00444.x (1999).

2 Martin-Trillo, M. & Cubas, P. TCP genes: a family snapshot ten years later. *Trends in plant science* **15**, 31-39, doi:10.1016/j.tplants.2009.11.003 (2010).

3 Riechmann, J. L. *et al.* *Arabidopsis* Transcription Factors: Genome-Wide Comparative Analysis among Eukaryotes. *Science* **290**, 2105-2110, doi:10.2307/3081600 (2000).

4 Yao, X., Ma, H., Wang, J. & Zhang, D. Genome-Wide Comparative Analysis and Expression Pattern of TCP Gene Families in *Arabidopsis* thaliana and Oryza sativa. *Journal of integrative plant biology* **49**, 885-897, doi:10.1111/j.1744-7909.2007.00509.x (2007).

5 Navaud, O., Dabos, P., Carnus, E., Tremousaygue, D. & Herve, C. TCP transcription factors predate the emergence of land plants. *Journal of molecular evolution* **65**, 23-33, doi:10.1007/s00239-006-0174-z (2007).

6 Pagnussat, G. C. *et al.* Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. *Development* **132**, 603-614, doi:10.1242/dev.01595 (2005).

7 Sarvepalli, K. & Nath, U. Hyper-activation of the TCP4 transcription factor in *Arabidopsis* thaliana accelerates multiple aspects of plant maturation. *The Plant journal : for cell and molecular biology* **67**, 595-607, doi:10.1111/j.1365-313X.2011.04616.x (2011).

8 Takeda, T. *et al.* RNA interference of the *Arabidopsis* putative transcription factor TCP16 gene results in abortion of early pollen development. *Plant molecular biology* **61**, 165-177, doi:10.1007/s11103-006-6265-9 (2006).

9 Giraud, E. *et al.* TCP transcription factors link the regulation of genes encoding mitochondrial proteins with the circadian clock in *Arabidopsis* thaliana. *The Plant cell* **22**, 3921-3934, doi:10.1105/tpc.110.074518 (2010).

10 Hammani, K. *et al.* An *Arabidopsis* dual-localized pentatricopeptide repeat protein interacts with nuclear proteins involved in gene expression regulation. *The Plant cell* **23**, 730-740, doi:10.1105/tpc.110.081638 (2011).

11 Takeda, T. *et al.* The OsTB1 gene negatively regulates lateral branching in rice. *The Plant journal : for cell and molecular biology* **33**, 513-520 (2003).

12 Aguilar-Martinez, J. A. & Sinha, N. Analysis of the role of *Arabidopsis* class I TCP genes AtTCP7, AtTCP8, AtTCP22, and AtTCP23 in leaf development. *Frontiers in plant science* **4**, 406, doi:10.3389/fpls.2013.00406 (2013).

13 Wang, M. Y. *et al.* The cotton transcription factor TCP14 functions in auxin-mediated epidermal cell differentiation and elongation. *Plant physiology* **162**, 1669-1680, doi:10.1104/pp.113.215673 (2013).

14 Hao, J. *et al.* GbTCP, a cotton TCP transcription factor, confers fibre elongation and root hair development by a complex regulating system. *Journal of experimental botany* **63**, 6267-6281, doi:10.1093/jxb/ers278 (2012).

15 Ma, J. *et al.* Expression profiles of miRNAs in Gossypium raimondii. *Journal of Zhejiang University. Science. B* **16**, 296-303, doi:10.1631/jzus.B1400277 (2015).

16 Paterson, A. H. *et al.* Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature* **492**, 423-427, doi:10.1038/nature11798 (2012).
17 Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**, 4673-4680 (1994).
18 de Castro, E. *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic acids research* **34**, W362-365, doi:10.1093/nar/gkl124 (2006).
19 Ma, J. *et al.* Genome-wide identification and expression analysis of TCP transcription factors in Gossypium raimondii. *Scientific reports* **4**, 6645, doi:10.1038/srep06645 (2014).
20 Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic acids research* **25**, 4876-4882 (1997).
21 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* **30**, 2725-2729, doi:10.1093/molbev/mst197 (2013).
22 Zhou, T. *et al.* Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. *Mol Genet Genomics* **271**, 402-415, doi:10.1007/s00438-004-0990-z (2004).
23 Yang, S., Zhang, X., Yue, J. X., Tian, D. & Chen, J. Q. Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol Genet Genomics* **280**, 187-198, doi:10.1007/s00438-008-0355-0 (2008).
24 Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452, doi:10.1093/bioinformatics/btp187 (2009).
25 Wang, K. *et al.* The draft genome of a diploid cotton Gossypium raimondii. *Nature genetics* **44**, 1098-1103, doi:10.1038/ng.2371 (2012).
26 Guo, A. Y., Zhu, Q. H., Chen, X. & Luo, J. C. [GSDS: a gene structure display server]. *Yi chuan = Hereditas / Zhongguo yi chuan xue hui bian ji* **29**, 1023-1026 (2007).
27 Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research* **34**, W369-373, doi:10.1093/nar/gkl198 (2006).
28 Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic acids research* **33**, W116-120, doi:10.1093/nar/gki442 (2005).
29 Li, F. *et al.* Genome sequence of the cultivated cotton Gossypium arboreum. *Nature genetics* **46**, 567-572, doi:10.1038/ng.2987 (2014).
30 Tatematsu, K., Nakabayashi, K., Kamiya, Y. & Nambara, E. Transcription factor AtTCP14 regulates embryonic growth potential during seed germination in *Arabidopsis* thaliana. *The Plant journal : for cell and molecular biology* **53**, 42-52, doi:10.1111/j.1365-313X.2007.03308.x (2008).
31 Rueda-Romero, P., Barrero-Sicilia, C., Gomez-Cadenas, A., Carbonero, P. & Onate-Sanchez, L. *Arabidopsis* thaliana DOF6 negatively affects germination in non-after-

ripened seeds and interacts with TCP14. *Journal of experimental botany* **63**, 1937-1949, doi:10.1093/jxb/err388 (2012).
32 Kieffer, M., Master, V., Waites, R. & Davies, B. TCP14 and TCP15 affect internode length and leaf shape in *Arabidopsis*. *The Plant journal : for cell and molecular biology* **68**, 147-158, doi:10.1111/j.1365-313X.2011.04674.x (2011).
33 Guo, Z. *et al.* TCP1 modulates brassinosteroid biosynthesis by regulating the expression of the key biosynthetic gene DWARF4 in *Arabidopsis* thaliana. *The Plant cell* **22**, 1161-1173, doi:10.1105/tpc.109.069203 (2010).
34 Koyama, T., Ohme-Takagi, M. & Sato, F. Generation of serrated and wavy petals by inhibition of the activity of TCP transcription factors in *Arabidopsis* thaliana. *Plant signaling & behavior* **6**, 697-699 (2011).
35 *Arabidopsis* Genome, I. Analysis of the genome sequence of the flowering plant *Arabidopsis* thaliana. *Nature* **408**, 796-815, doi:10.1038/35048692 (2000).
36 Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant cell* **16**, 1667-1678, doi:10.1105/tpc.021345 (2004).
37 Zhang, J. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18**, 292-298, doi:http://dx.doi.org/10.1016/S0169-5347(03)00033-8 (2003).
38 Flagel, L. E. & Wendel, J. F. Gene duplication and evolutionary novelty in plants. *The New phytologist* **183**, 557-564, doi:10.1111/j.1469-8137.2009.02923.x (2009).
39 Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nature reviews. Genetics* **10**, 725-732, doi:10.1038/nrg2600 (2009).
40 Lecharny, A., Boudet, N., Gy, I., Aubourg, S. & Kreis, M. Introns in, introns out in plant gene families: a genomic approach of the dynamics of gene structure. *Journal of structural and functional genomics* **3**, 111-116 (2003).

Table 2.1 TCP gene family in *G. arboreum*

Gene Name	Gene symbol	Length(aa)	MW(Da)	pI	Chr. Location
GaTCP1	Cotton_A_09911	397	43545.24	9.21	chr3:16012758:16014321
GaTCP2	Cotton_A_26168	410	44917.18	6.77	chr10:15878745:15879977
GaTCP3	Cotton_A_23161	409	44273.72	6.78	chr13:94068971:94070200
GaTCP4	Cotton_A_22289	443	48760.48	6.06	chr10:78860246:78861667
GaTCP5	Cotton_A_31971	325	36033.79	5.8	chr1:70185179:70186156
GaTCP6	Cotton_A_23025	250	26502.48	9.58	chr6:20222965:20223732
GaTCP7a	Cotton_A_08973	258	26739.99	9.72	chr5:10164781:10165557
GaTCP7b	Cotton_A_14593	243	25308.51	9.99	chr6:109156974:109157705
GaTCP8	Cotton_A_24144	486	50969.89	7.36	chr5:48534877:48536337
GaTCP9a	Cotton_A_10947	338	35365.29	9.08	chr4:104148560:104149576
GaTCP9b	Cotton_A_14431	385	41254.45	8.95	chr7:47526308:47527465
GaTCP10	Cotton_A_20110	448	48746.25	6.6	chr7:116121419:116122765
GaTCP11	Cotton_A_24059	200	21687.42	8.32	chr10:109535667:109536269
GaTCP12	Cotton_A_37122	501	55859.83	6.82	chr7:89657212:89658717
GaTCP13a	Cotton_A_27227	309	34245.51	8.71	chr12:56165133:56166062
GaTCP13b	Cotton_A_14726	285	31976.83	7.94	chr11:102879533:102880390
GaTCP14a	Cotton_A_09220	395	42218.14	6.91	chr13:44785981:44787168
GaTCP14b	Cotton_A_02703	418	44468.80	8.60	chr8:96754543:96755799
GaTCP14c	Cotton_A_27685	406	43980.41	6.88	chr6:58440407:58441627
GaTCP15a	Cotton_A_06142	342	37377.38	8.53	chr6:68046188:68047216
GaTCP15b	Cotton_A_33342	365	39684.16	9.54	chr2:18664013:18665110
GaTCP16	Cotton_A_10509	196	21078.67	8.56	chr8:103306077:103306667
GaTCP17	Cotton_A_19125	266	30312.08	7.78	chr1:83153095:83153967
GaTCP18a	Cotton_A_07573	329	37748.89	9.02	chr11:53845238:53846327
GaTCP18b	Cotton_A_01394	367	41500.33	9.08	chr6:120385396:120386499
GaTCP19a	Cotton_A_21588	341	36882.22	6.26	chr13:4529921:4530946
GaTCP19b	Cotton_A_09964	335	35753.32	9.42	chr3:14665842:14666849
GaTCP20a	Cotton_A_40823	300	32008.71	7.93	chr3:18568141:18569043
GaTCP20b	Cotton_A_07501	298	31710.22	7.28	chr9:49546327:49547223
GaTCP20c	Cotton_A_39272	298	31420.10	9.64	chr11:43972643:43973539
GaTCP20d	Cotton_A_22689	306	32794.32	9.27	chr1:99325941:99326861
GaTCP21	Cotton_A_26482	255	26370.49	9.66	chr13:2155968:2156735
GaTCP22	Cotton_A_27060	553	58300.67	6.73	chr2:54926878:54928539
GaTCP23	Cotton_A_03998	418	44401.83	6.72	chr10:80131534:80132790
GaTCP24	Cotton_A_02913	463	50231.92	7.01	chr9:73463153:73464544
GaTCP25	Cotton_A_37650	431	47737.38	6.47	chr12:125507830:125509978

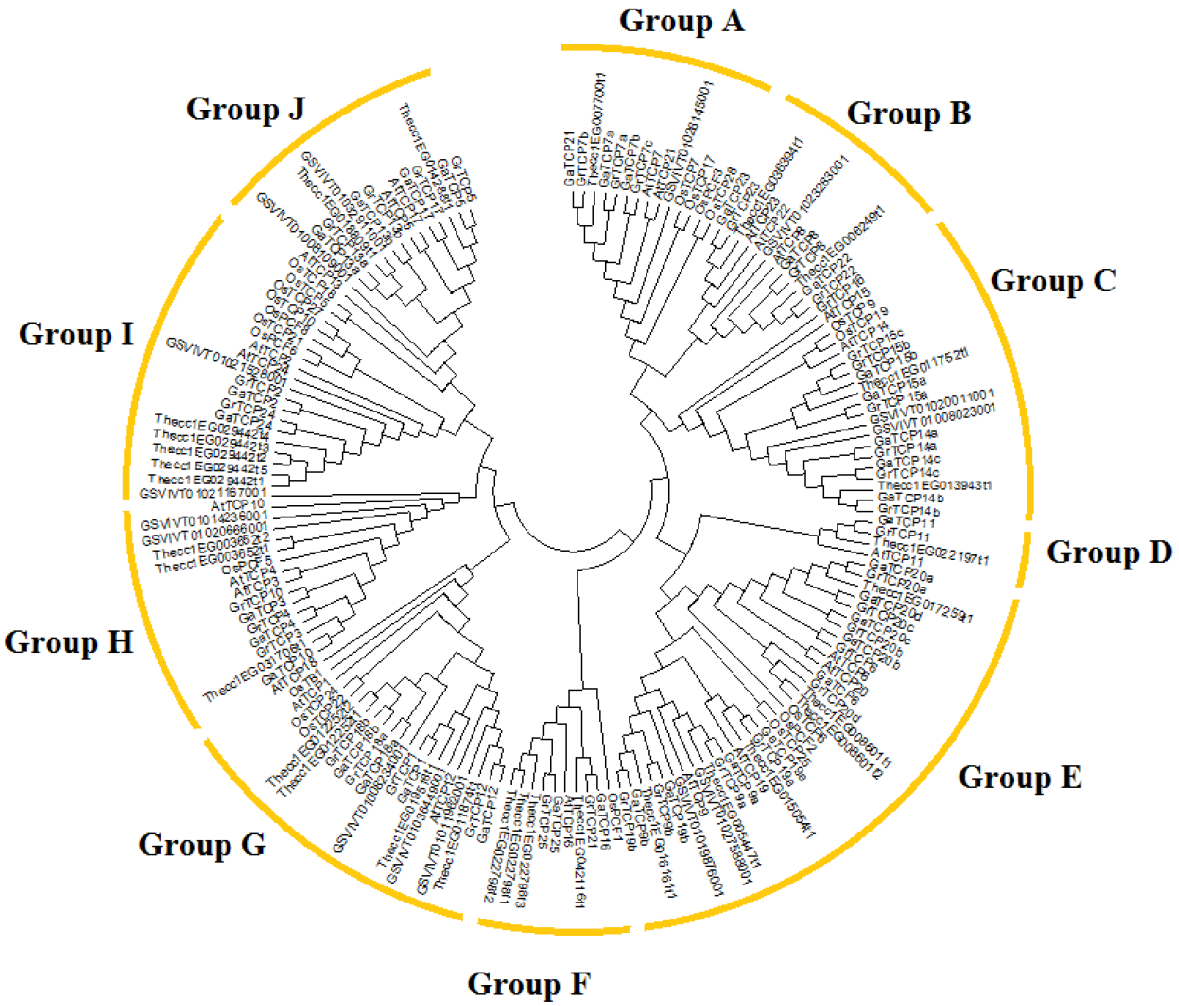


Figure 2.1. Phylogenetic tree of TCP proteins from *Gossypium arboreum*, *G. raimondii*, *Theobroma cacao*, *Vitis vinifera*, *Arabidopsis thaliana* and *Oryza sativa*. The phylogenetic tree was generated using the Neighbor-Joining (NJ) method implemented in the MEGA 6.0 software with JTT model and pairwise gap deletion option. The bootstrap analysis was conducted with 1000 iterations.

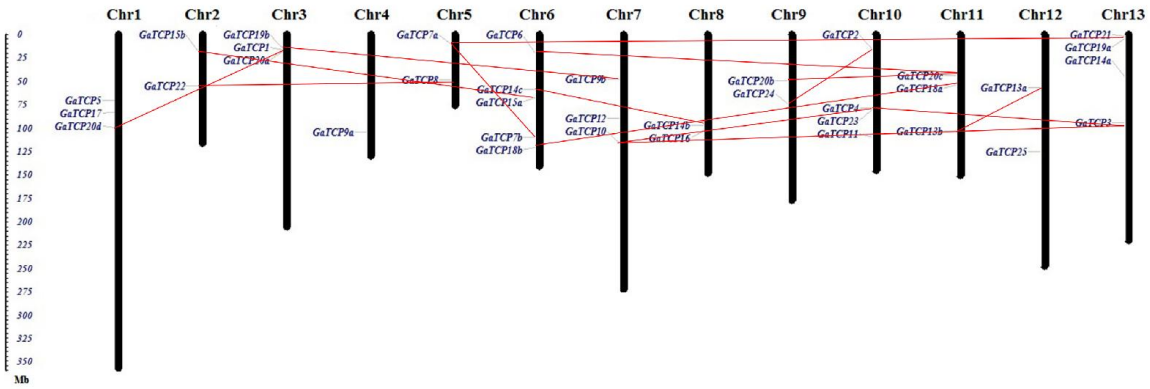


Figure 2.2. Chromosomal location and gene duplication status of TCP genes from *Gossypium arboreum* on 13 chromosomes. The scale represents megabases (Mb). The chromosome numbers are indicated above each vertical bar. The red lines connect the duplicated gene pairs.

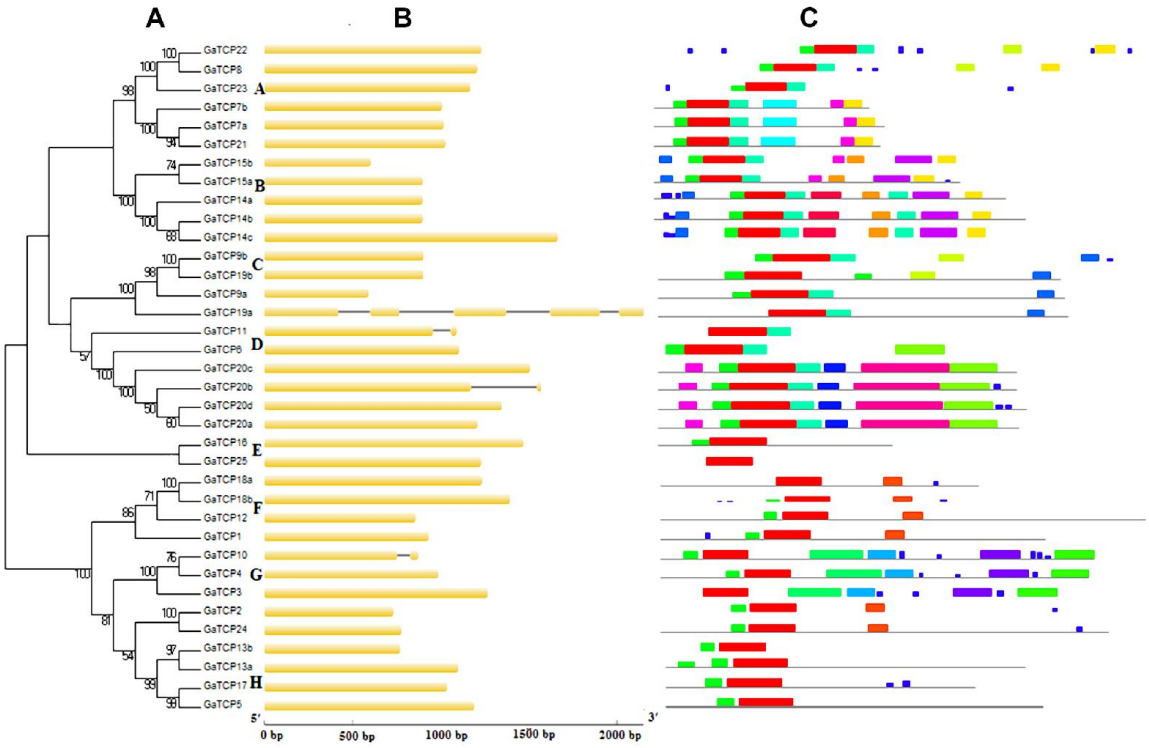


Figure 2.3. Phylogenetic analysis, exon/intron organization and motif composition s of *Gossypium arboreum* TCP genes. A. The phylogenetic tree was generated using the Neighbor-Joining (NJ) method implemented in the MEGA 6.0 software with JTT model and pairwise gap deletion option. The bootstrap analysis was conducted with 1000 iterations. B. the exon/intron distribution of *G. arboreum* TCP genes. Exons and introns are represented by green boxes and black lines, respectively. C. the motif compositions of *G. arboreum* TCP genes. Each color represents a specific motif.

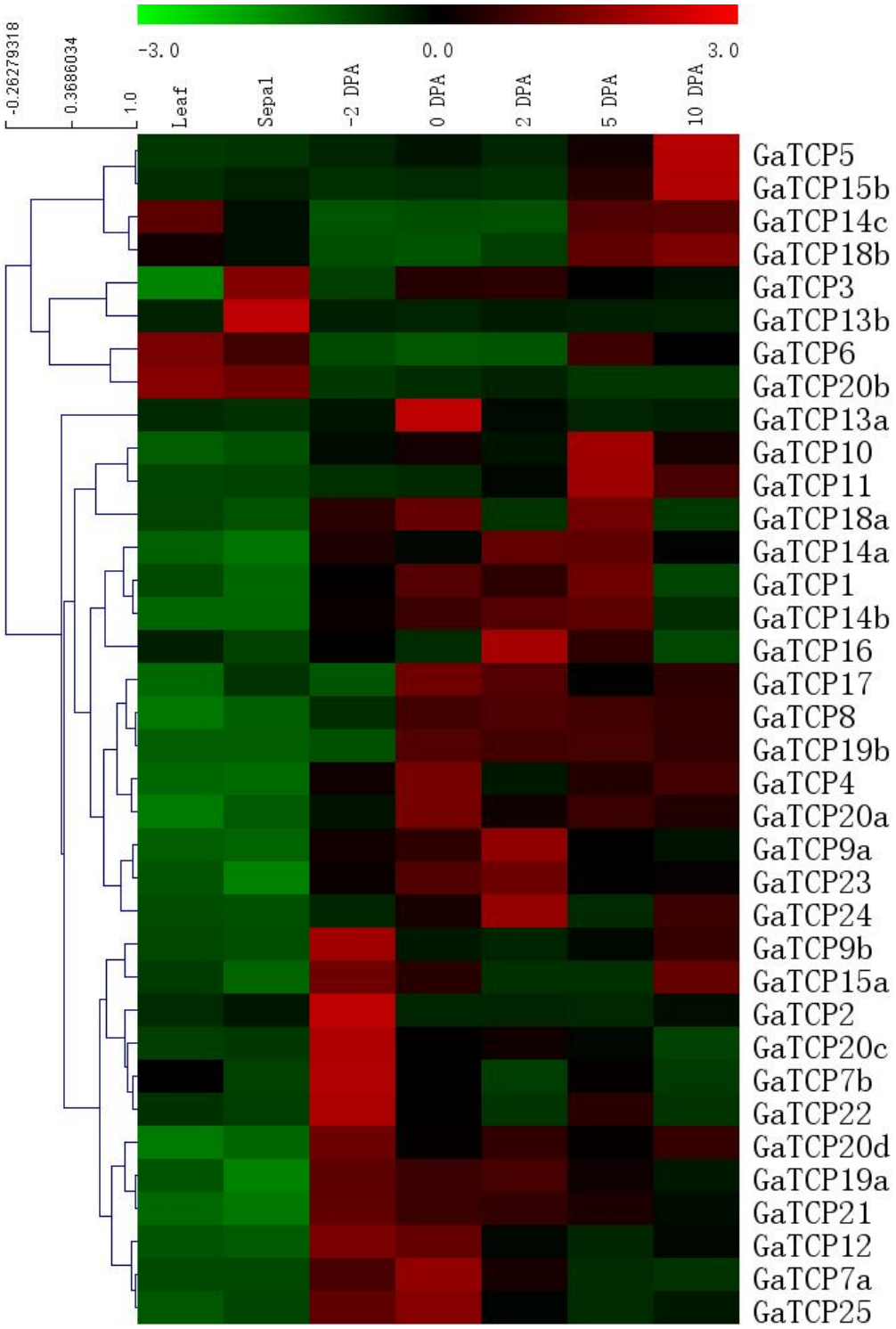


Figure 2.4. Expression profiles of *G. arboreum* TCP genes in different tissues and at different fiber developmental stage. The expression levels are represented by the color bar.

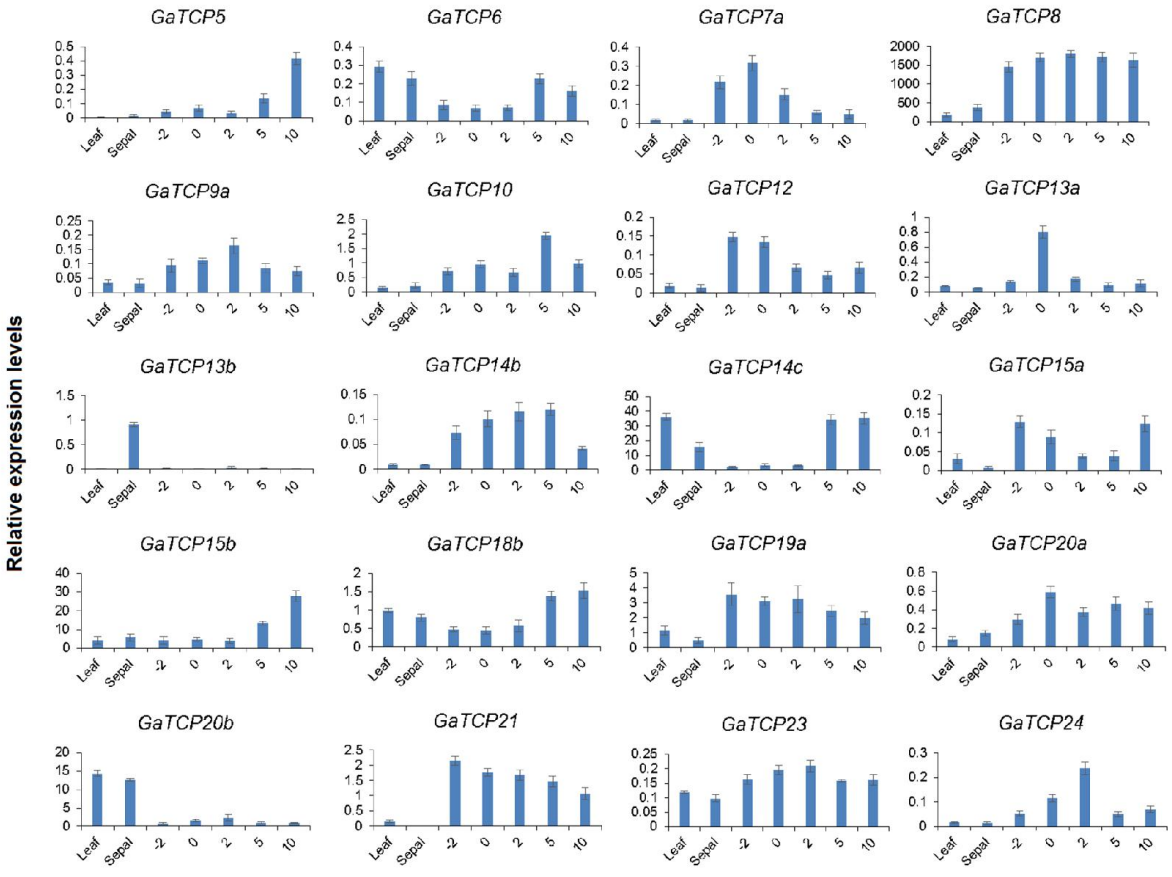


Figure 2.5. Expression profiles of 20 *G. arboreum* TCP genes in different tissues and at different fiber developmental stage. The X-axis represents different tissues or developmental stages while the Y-axis represents relative expression levels against the reference gene *SAD1*. Error bars are drawn based on standard deviation for three replicates.